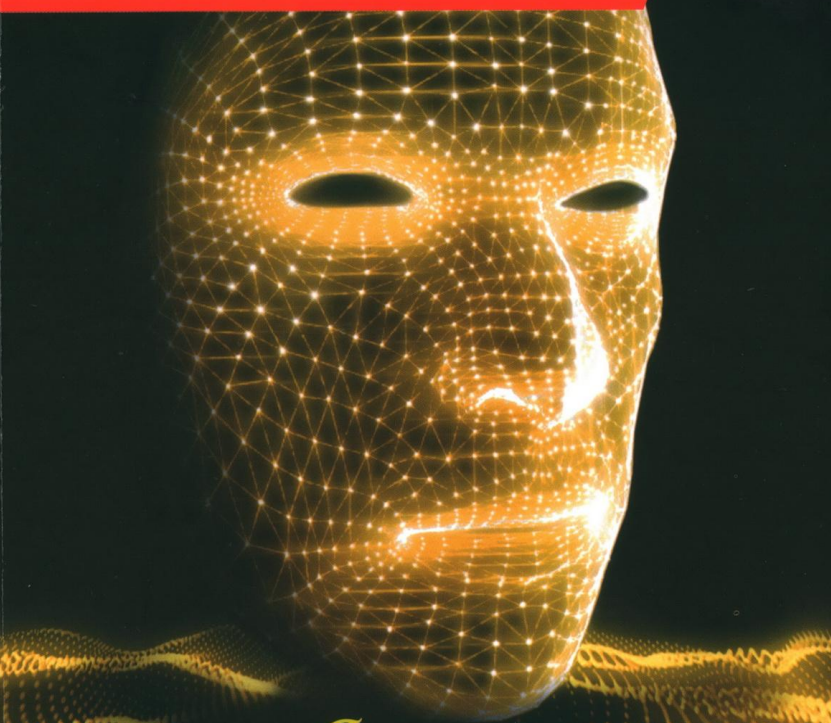
VERSANT

JOHN BROCKMAN (ed.)

MINȚI POSIBILE

Douăzeci și cinci de perspective
asupra inteligenței artificiale



VELLANT
CĂRȚI ÎNSEMNATE.

JOHN BROCKMAN

este un impresar cultural a cărui carieră s-a intersectat cu domenii precum arta de avangardă, știința, publishing-ul, dezvoltarea de software sau Internetul. Este fondatorul platformei edge.org și editorul seriei de cărți Edge Question, care include titluri precum This Idea is Brilliant, Know This, This Idea Must Die, This Explains Everything, This Will Make You Smarter.

JOHN
BROCKMAN
(EDITOR)

Minți posibile

VERSANT

MINȚI POSIBILE

Traducere din limba engleză
de George Chiriță

VELLANT
CĂRȚI ÎNSEMNATE.

Redactor: Doinel Tronaru
Corector: Dușa Udrea-Boborel
Copertă: Daniel Breanova
DTP: Răzvan Nasea
Coordonare editorială: Doru Someșan
Consultant editorial: Simona Radu-Răuță
Concept și identitate vizuală: RUSU+BORȚUN

Descrierea CIP a Bibliotecii Naționale a României
Minți posibile : douăzeci și cinci de perspective asupra inteligenței artificiale / ed.: John Brockman ; trad.: George Chiriță. - București : Editura Vellant, 2019
ISBN 978-606-980-067-6

I. Brockman, John (ed.)
II. Chiriță, George (trad.)

004

Titlu original: *Possible Minds: Twenty-Five Ways of Looking at AI*
Copyright © 2019 by John Brockman. All rights reserved.

© 2019 Editura Vellant

Editura Vellant
Splaiul Independenței 319
București, Sector 6
www.vellant.ro

Toate drepturile asupra versiunii în limba română
aparțin Editurii Vellant.

Pentru Einstein, Gertrude Stein,
Wittgenstein și Frankenstein

Mulțumiri

Multe mulțumiri lui Scott Moyers de la Penguin Press, pentru exuberanța sa editorială, și agentului meu, Max Brockman, pentru încurajările sale continue. Mulțumiri speciale, din nou, Sarei Lippincott, pentru atenția acordată acestui manuscris.

Cuprins

Introducere
Despre promisiunile și pericolele inteligenței artificiale 11

Capitolul 1. Seth Lloyd:
Greșit, dar mai relevant decât oricând 25

Capitolul 2. Judea Pearl:
Limitările mașinilor de învățare opace 39

Capitolul 3. Stuart Russell:
Scopul scris în mașină 46

Capitolul 4. George Dyson:
A treia lege 62

Capitolul 5. Daniel C. Dennett:
Ce putem face? 70

Capitolul 6. Rodney Brooks:
Debandada inumană în care ne-au băgat mașinile 84

Capitolul 7. Frank Wilczek:
Unitatea inteligenței 95

Capitolul 8. Max Tegmark:
Să aspirăm la mai mult decât doar să nu devenim depășiți 107

Capitolul 9. Jaan Tallinn:
Mesaje disidente 120

Capitolul 10. Steven Pinker:
Profeția tehnologică și puterea cauzală subapreciată a ideilor 133

Capitolul 11. David Deutsch:
Dincolo de recompensă și pedeapsă 147

Capitolul 12. Tom Griffiths:
Folosirea artificială a ființelor umane 161

Capitolul 13. Anca Drăgan:
Integrarea omului în ecuația inteligenței artificiale 171

Capitolul 14. Chris Anderson:
Coborârea pe panta unui gradient 181

Capitolul 15. David Kaiser:
„Informația" pentru Wiener, pentru Shannon și pentru noi 191

Capitolul 16. Neil Gershenfeld:
Scalarea 201

Capitolul 17. W. Daniel Hills:
Primele inteligențe automate 212

Capitolul 18. Venki Ramakrishnan:
Vor deveni computerele stăpânii noștri 224

Capitolul 19. Alex „Sandy" Pentland:
Strategia umană 237

Capitolul 20. Hans Ulrich Obrist:
Transformarea invizibilului în vizibil: arta față în față cu inteligența artificială 253

Capitolul 21. Alison Gopnik:
Inteligențele artificiale versus copiii de patru ani 267

Capitolul 22. Peter Galison:
Algoritmicienii visează la obiectivitate 280

Capitolul 23. George M. Church:
Drepturile mașinilor 290

Capitolul 24. Caroline A. Jones:
Utilizarea artistică a ființelor cibernetice 305

Capitolul 25. Stephen Wolfram:
Inteligența artificială și viitorul civilizației 318

Introducere
Despre promisiunile și pericolele inteligenței artificiale

Inteligența artificială este povestea zilei de astăzi, povestea din spatele tuturor celorlalte povești, este a doua venire a lui Iisus și Apocalipsa în același timp: inteligența artificială bună și inteligența artificială malefică. Această carte a reieșit în urma unor conversații în desfășurare cu câțiva gânditori importanți, atât din lumea inteligenței artificiale, cât și din afara acesteia, despre ce este și ce înseamnă inteligența artificială. Numită Proiectul Minți Posibile, această conversație a început în luna septembrie 2016, printr-o întâlnire la Grace Mayflower Inn & Spa din Washington, Connecticut, cu câțiva dintre cei care au contribuit la această carte.

Ce a reieșit foarte repede din acea primă întâlnire a fost că entuziasmul și teama din cultura mai largă ce înconjoară inteligența artificială acum au un corespondent în modul în care ideile lui Norbert Wiener cu privire la „cibernetică" și-au făcut loc în cultură, în special în anii 1960, pe măsură ce artiștii au început să încorporeze gândirea despre noile tehnologii în activitatea lor. Am fost martor apropiat la impactul acelor idei; și nu exagerez când spun că ele mi-au determinat traiectoria vieții. Odată cu sosirea erei digitale, la începutul anilor 1970, oamenii au încetat să mai vorbească despre Wiener, dar astăzi această idee a cibernetică a fost adoptată la scară atât de largă, încât este internalizată

până la punctul în care nu mai are nevoie de nume. Este peste tot, este în aer și este un loc potrivit de unde să începem.

Tehnologii noi = percepții noi

Înainte de inteligența artificială, a existat cibernetica – ideea controlului automat, care se reglează singur, expusă în textul de referință al lui Norbert Wiener din 1948. Pot să-mi datez propria mea expunere serioasă la acest test în 1966, când compozitorul John Cage m-a invitat pe mine și pe încă patru sau cinci artiști tineri să participăm împreună la o serie de serate – un seminar în desfășurare despre mass-media, comunicații, artă, muzică și filosofie, care se concentra pe interesul lui Cage în ceea ce privește ideile lui Wiener, Claude Shannon și Marshall McLuhan, toți având greutate în cercurile artistice din New York în care mă învârteam atunci. În special Cage preluase ideea lui McLuhan conform căreia, prin inventarea tehnologiilor electronice, am externalizat sistemul nostru nervos central – adică mințile noastre – și acum trebuia să presupunem că „există o singură minte, cea pe care o împărtășim cu toții".

Ideile de acest fel începeau să fie interesante pentru artiștii cu care lucram la New York și la Film-Makers' Cinematheque, unde eram director de programe artistice pentru o serie de producții multimedia denumită New Cinema 1 (cunoscută și sub numele Expanded Cinema Festival), sub auspiciile regizorului avangardist de film și impresarului Jonas Mekas. Printre ei se numărau: Claes Oldenburg, Robert Rauschenberg, Andy Warhol și Robert Whitman; artiștii cinetici: Charlotte Moorman și Nam June Paik; artiștii momentului Allan Kaprow și Carolee Schneemann; dansatoarea Trisha Brown; regizorii: Jack Smith, Stan Vanderbeek, Ed Emshwiller și frații Kuchar; dramaturgul de avangardă Ken Dewey; poetul Gern Stern și grupul USCO; muzicienii minimaliști La Monte Young și Terry Riley; și, prin Warhol, grupul muzical The Velvet Underground. Mulți dintre acești oameni citeau Wiener și cibernetica era în aer. La una dintre aceste

serate, Cage a băgat mâna în valiza sa și a scos un exemplar al cărții *Cibernetica*, mi l-a înmânat și a spus „Asta e pentru tine".

În timpul festivalului, am primit un apel telefonic neașteptat de la colegul lui Wiener, Arthur K. Solomon, coordonatorul masteratelor și doctoratelor de la Harvard în biofizică. Wiener murise cu un an înainte și ceilalți colegi apropiați ai lui Wiener și Solomon de la MIT și Harvard citiseră despre Expanded Cinema Festival în *New York Times*, fiind intrigați de legătura cu cartea lui Wiener. Solomon m-a invitat să aduc câțiva artiști la Cambridge pentru a mă întâlni cu el și cu un grup ce-i includea și pe cercetătorul în domeniul comunicațiilor senzoriale de la MIT, Walter Rosenblith, pe matematicianul de la Harvard Anthony Oettinger și pe un inginer MIT Harold „Doc" Edgerton, inventatorul luminii stroboscopice.

Ca multe alte situații în care „arta întâlnește știința" în care am fost implicat de atunci, evenimentul de două zile era un rateu informat: nave care se petrec în noapte. Dar m-am înhămat la acest proiect și evenimentul a fost important din câteva puncte de vedere interesante – unul fiind faptul că ne-au dus să vedem Computerul. Computerele erau o raritate pe atunci; cel puțin niciunul dintre noi, cei care eram în vizită, nu văzuserăm vreodată unul. Am fost conduși într-un spațiu mare din campusul MIT, în mijlocul căruia se afla o „cameră rece" ridicată de la sol și cu pereți de sticlă, în care tehnicieni cu halate, eșarfe și mănuși albe erau ocupați cu colaționarea unor cartele perforate care ieșeau dintr-o mașină enormă. Când m-am apropiat, aburul respirației mele a aburit fereastra din camera rece. Ștergând fereastra, am văzut Computerul. M-am îndrăgostit.

Mai târziu, în toamna anului 1967, am mers în Menlo Park pentru a petrece puțin timp cu Stewart Brand, pe care l-am întâlnit în New York, în 1965, când era un membru-satelit al grupului de artiști USCO. Acum, cu soția sa Lois, matematiciană, pregătea pentru publicare prima ediție a revistei *Whole Earth Catalog*. În timp ce Lois și echipa făceau munca grea pentru ultimele mecanici ale *WEC*, eu și Stewart am stat împreună într-un colț timp

de două zile, citind, subliniind și adnotând exemplarul din *Cibernetica* pe care mi-l dăduse Cage cu un an înainte și dezbătând ideile lui Wiener.

Inspirat de acest set de idei, am început să dezvolt o temă, un fel de mantră, care mi-a ghidat încercările de atunci: „tehnologii noi înseamnă percepții noi". Inspirat de teoreticianul comunicării Marshall McLuhan, designerul-arhitect Buckminster Fuller, futuristul John McHale și antropologii culturali Edward T. „Ned" Hall și Edmund Carpenter, am început să citesc cu nesaț despre teoria informației, cibernetică și teoria sistemelor. McLuhan mi-a sugerat să citesc cartea biologului J.Z. Young *Doubt and Certainty in Science*, în care spune că noi creăm instrumente și ne modelăm prin utilizarea acestora. Celălalt text pe care l-a recomandat era lucrarea din 1949 a lui Warren Weaver și Claude Shannon intitulată *Contribuții recente la teoria matematică a comunicării*, care începe așa: „Cuvântul *comunicare* va fi utilizat aici cu sens foarte larg, pentru a include toate procedurile prin care o minte o poate afecta pe alta. Bineînțeles, acest lucru nu implică numai limbajul scris și oral, ci și muzica, pictura, teatrul, baletul și, de fapt, tot comportamentul uman".

Cine ar fi știut că în două decenii de la acel moment vom începe să recunoaștem creierul ca pe un computer? Și, în următoarele două decenii, pe măsură ce am integrat computerele noastre pe internet, că va urma să realizăm despre creierul nostru că nu este un computer, ci o rețea de computere? Cu siguranță nu Wiener, specialist în circuitele cu *feedback* analogice concepute pentru a controla mașinile, nici artiștii, cu atât mai puțin eu.

„Trebuie să încetăm să sărutăm biciul ce ne biciuiește"

La doi ani după *Cibernetica*, în 1950, Norbert Wiener a publicat cartea *The Human Use of Human Beings**, o poveste mai

* Pe parcursul cărții, vom folosi titlul cărții în limba engleză, deoarece această carte nu a fost tradusă în limba română. O traducere aproximativă a titlului ar fi *Folosirea omului de către om* (n.t.).

profundă, în care își exprima îngrijorările despre exploatarea comercială scăpată de sub control și alte consecințe neprevăzute ale noilor tehnologii de control. Nu am citit *The Human Use of Human Beings* până în primăvara anului 2016, când am găsit un exemplar, ediție princeps, care stătea în biblioteca mea lângă *Cibernetica*. Ce m-a șocat a fost să realizez cât de profetic a putut să fie Wiener în 1950 despre ce se întâmplă în zilele noastre. Deși prima ediție a fost un *bestseller* major – și, într-adevăr, a inițiat o conversație importantă –, sub presiunea colegilor săi de breaslă, Wiener a scos o ediție revizuită și mai temperată în 1954, din care capitolul de concluzii inițial, „Vocile rigidității", lipsește.

Istoricul științei George Dyson evidențiază faptul că, în această ediție princeps de mult uitată, Wiener a preconizat posibilitatea „unui fascism nou și amenințător, dependent de *machine à gouverner*":

> Nicio elită nu a scăpat de critica sa, de la marxiști și iezuiți („tot catolicismul este esențialmente o religie totalitară") până la FBI („marii noștri prinți negustori au analizat tehnica de propagandă a rușilor și au constatat că este bună") și mogulii care își oferă sprijinul în „a face capitalismul american și a cincea libertate a comerciantului supreme în lume". Oamenii de știință... au fost supuși aceleiași cercetări meticuloase la care a fost supusă și Biserica: „Într-adevăr, directorii marilor laboratoare seamănă foarte mult cu episcopii, cu asocierile lor cu persoane influente din toate palierele vieții și pericolele la care se supun din cauza păcatelor carnale, a mândriei și a lăcomiei de putere".

Ieremiada nu a mers bine pentru Wiener. În cuvintele lui Dyson:

> Acele semnale de alarmă au fost ignorate la vremea aceea nu pentru că Wiener se înșela în ceea ce privește tehnica de calcul digitală, ci pentru că existau amenințări

mai mari când și-a terminat manuscrisul în toamna anului 1949. Wiener nu avea nimic împotriva tehnicii de calcul digitale, dar se opunea cu tărie armelor nucleare și a refuzat să se alăture celor care construiau computere digitale pentru a avansa cercetările pentru bomba cu hidrogen, care era de o mie de ori mai puternică.

Deoarece prima ediție a cărții *The Human Use of Human Beings* este acum ieșită de pe piață, nu mai avem acces la *cri-de-coeur*-ul lui Wiener, mai relevant astăzi decât atunci când l-a scris, acum șaizeci și opt de ani: „Trebuie să încetăm să sărutăm biciul ce ne biciuiește".

Mintea, gândirea, inteligența

Printre motivele pentru care nu auzim prea multe despre cibernetică în zilele noastre, două sunt centrale: în primul rând, deși *The Human Use of Human Beings* a fost considerată o carte importantă la vremea sa, era contrară aspirațiilor multor colegi ai lui Wiener, printre care și John von Neumann și Claude Shannon, care erau interesați de comercializarea noilor tehnologii. În al doilea rând, pionierul computerelor, John McCarthy, nu-l agrea pe Wiener și refuza să folosească termenul acestuia de „cibernetică". McCarthy, la rândul său, a născocit sintagma de „inteligență artificială" și a devenit un „părinte fondator" al acestui domeniu.

După cum îmi explica Judea Pearl, care, în anii 1980, a introdus o abordare nouă a inteligenței artificiale, denumită rețelele bayesiene:

> Ceea ce a creat Wiener a fost entuziasmul de a crede că într-o bună zi vom construi o mașină inteligentă. Nu era un informatician. Vorbea despre *feedback*, despre comunicare, despre analogic. Metafora prin care se exprima era un circuit cu feedback, domeniu în care era expert. Când a

debutat era digitală, la începutul anilor 1960, oamenii voiau să vorbească despre programare, despre coduri, despre funcții computaționale, despre memoria pe termen scurt, memoria pe termen lung – metafore informatice care musteau de sens. Wiener nu făcea parte din acest val și nu a ajuns la noua generație ce se născuse cu ideile sale. Metaforele sale erau prea vechi, *passé*. Erau deja disponibile mijloace noi care erau pregătite să captiveze imaginația umană. Pe la începutul anilor 1970, oamenii nu mai vorbeau despre Wiener.

Un factor critic care lipsea din viziunea lui Wiener era elementul cognitiv: mintea, gândirea și inteligența. Încă din 1942, la prima întâlnire dintr-o serie de întâlniri interdisciplinare despre controlul sistemelor complexe care vor fi cunoscute sub numele de Conferințele Macy, cercetători de top argumentau includerea elementului cognitiv în conversație. În timp ce von Neumann, Shannon și Wiener erau concentrați pe sistemele de control și comunicare ale sistemelor observate, Warren McCullough voia să includă și mintea. S-a orientat către antropologii culturali Gregory Bateson și Margaret Mead, pentru a face legătura cu științele sociale. Bateson, în special, vorbea din ce în ce mai mult despre tipare și procese sau despre „tiparul care conectează". A solicitat un nou tip de ecologie a sistemelor, în care organismele și mediul în care trăiesc sunt unul și același și ar trebui considerate un singur circuit. La începutul anilor 1970, cibernetica sistemelor observate – cibernetica de rangul întâi – s-a deplasat către cibernetica sistemelor observatoare – cibernetica de rangul al doilea sau „cibernetica ciberneticii", termen introdus de Heinz von Foerster, care s-a alăturat conferințelor Macy la jumătatea anilor 1950 și a condus o nouă mișcare.

Cibernetica, în loc să dispară, se metaboliza în *tot*, așa că nu o mai vedeam ca pe o nouă disciplină separată și distinctă. Și acolo a rămas, ascunsă în văzul tuturor.

„Sceneta actorilor principali"

Propriile mele scrieri pe această temă la momentul respectiv erau pe radarul oamenilor care se ocupau de cibernetica de rangul al doilea, printre care Heinz von Foerster, dar și John Lilly și Alan Watts, care erau coorganizatorii a ceva numit Conferințele AUM, acronim pentru „the American University of Masters", care a avut loc în Big Sur, în 1973, o adunare de filosofi, psihologi și oameni de știință, fiecare personalitate trebuind să țină o prelegere despre propria sa activitate în termenii relației acesteia cu ideile matematicianului britanic G. Spencer-Brown, prezentate în cartea sa *Laws of Form*. Am fost puțin nedumerit când am primit invitația – o invitație foarte târzie –, despre care spuneau că se bazează pe interesul lor în ideile prezentate de mine într-o carte cu titlul *Afterwords*, idei care erau pe lungimea lor de undă. Am profitat bucuros de ocazie, motivul principal fiind că vorbitorul principal era nimeni altul decât Richard Feynman. Ador să-mi petrec timpul cu fizicieni, deoarece se gândesc la univers, adică la tot. Și niciun fizician nu avea o reputație de a fi mai articulat decât Feynman. Abia așteptam să-l cunosc. Am acceptat. Acestea fiind spuse, nu sunt un om de știință și nu m-am gândit niciodată să mă urc pe o scenă și să țin vreo „prelegere" de orice fel, cu atât mai puțin un comentariu despre o teorie matematică obscură în fața unui grup identificat ca fiind format dintre cei mai interesanți gânditori din lume. Abia după ce am sosit în Big Sur, am aflat motivul invitației mele întârziate. „Când va vorbi Feynman?" am întrebat eu la recepție. „A, nu ți-a spus Alan Watts? Richard este bolnav și a fost internat. Ești înlocuitorul lui. Și, apropo, care este titlul discursului tău inaugural?"

Am încercat să mă fac invizibil câteva zile. Alan Watts, dându-și seama că evit podiumul, m-a trezit într-o noapte, la trei dimineața, ciocănind la ușă. Am deschis ușa și l-am văzut stând în fața mea într-o robă de călugăr, cu gluga acoperindu-i o mare parte din față. Avea brațele întinse, o lanternă într-o mână și un whisky Magnum în cealaltă. „John", a spus el pe o voce groasă și cu un accent britanic

aristocratic, „ești un impostor. Și, John", a continuat „eu sunt un impostor. Dar, John, eu sunt un impostor *adevărat*!"

A doua zi, am ținut discursul, intitulat „Einstein, Gertrude Stein, Wittgenstein și Frankenstein". Einstein, revoluția din fizica secolului XX. Gertrude Stein, prima scriitoare care și-a integrat în lucrări ideea unui univers nedeterminat și discontinuu. Cuvintele nu reprezentau nici caracterul, nici activitatea: un trandafir este un trandafir este un trandafir, iar un univers este un univers este un univers. Wittgenstein: lumea ca limite ale limbajului. „Limitele limbajului meu înseamnă limitele lumii mele." Sfârșitul distincției dintre observator și observat. Frankenstein: cibernetica, inteligența artificială, robotica, toți eseiștii din acest volum.

Prelegerea a avut consecințe neprevăzute. Printre participanții la Conferințele AUM erau mai mulți autori de bestselleruri ajunse numărul unu în *New York Times*, însă nimeni de acolo nu avea un agent literar. Și mi-am dat seama că toată lumea scria un gen de carte nedenumit și nerecunoscut de editurile din New York. Deoarece aveam un masterat în administrarea afacerilor de la Columbia Business School și o serie de succese relative în afaceri, am fost hărțuit să devin agent, inițial pentru Gregory Bateson și John Lilly, ale căror cărți le-am vândut rapid și pentru sume care mi-au captat atenția, începându-mi astfel cariera de agent literar.

Nu am apucat să-l întâlnesc niciodată pe Richard Feynman.

Lungile ierni ale inteligenței artificiale

Această nouă carieră m-a pus în legătură apropiată cu majoritatea pionierilor din domeniu și, de-a lungul deceniilor, am plutit cu ei pe valuri de entuziasm și în văi de dezamăgire. La începutul anilor 1980, guvernul japonez a făcut un efort național de a avansa inteligența artificială. L-au numit a cincea generație; obiectivul lor era să schimbe arhitectura tehnicii de calcul trecând de „problema limitării a lui von Neumann" și creând un computer care lucrează masiv în paralel. Făcând acest

lucru, sperau să-și dezvolte economia și să devină o putere mondială dominantă în domeniu. În 1983, liderul consorțiului celei de-a Cincea Generații Japoneze a venit în New York pentru o întâlnire organizată de Heinz Pagels, președintele Academiei de Științe din New York. Aveam un loc la masă, alături de liderii primei generații Marvin Minsky și John McCarthy; de cei ai celei de-a doua generații, Edward Feigenbaum și Roger Schank; și de Joseph Traub, directorul Consorțiului Național de Supercomputere.

În 1981, cu ajutorul lui Heinz, am înființat The Reality Club (precursorul organizației nonprofit Edge.org), ale cărei întâlniri interdisciplinare inițiale aveau loc în sala de consiliu a Academiei de Științe din New York. Heinz lucra la cartea sa *The Dreams of Reason: The Computer and the Rise of the Science of Complexity*, pe care o considera a fi o agendă de cercetări pentru știință în anii 1990.

La întâlnirile Reality Club, am cunoscut doi cercetători tineri care urmau să joace roluri-cheie în revoluționarea informaticii. La MIT, la sfârșitul anilor 1970, Danny Hillis a dezvoltat algoritmii care au făcut posibil computerul masiv în paralel. În 1983, compania sa, Thinking Machines, a construit cel mai rapid supercomputer din lume utilizând arhitectura în paralel. „Mașina sa de conexiuni" reflecta îndeaproape modul în care funcționează mintea omenească. Seth Lloyd, de la Universitatea Rockefeller, desfășura o activitate foarte importantă în domeniile tehnicii de calcul cuantice și comunicațiilor cuantice, propunând inclusiv primul design fezabil din punct de vedere tehnologic pentru un computer cuantic.

Iar japonezii? Incursiunea lor în domeniul inteligenței artificiale a eșuat și a fost urmată de douăzeci de ani de creștere economică anemică. Însă cercetătorii importanți din SUA au luat acest program foarte în serios. Iar Feigenbaum, care era informaticianul cel mai cunoscut al momentului, a făcut echipă cu Pamela McCorduck pentru a scrie o carte despre aceste dezvoltări. *The Fifth Generation: Artificial Intelligence and Japan's Computer*

Challenge to the World a fost publicată în 1983. Aveam un nume de cod pentru proiect: „Vine! Vine!" Dar nu a venit; a plecat.

Din acel moment, am lucrat cu cercetători în aproape orice varietate a inteligenței artificiale și complexității, printre care: Rodney Brooks, Hans Moravec, John Archibald Wheeler, Benoit Mandelbrot, John Henry Holland, Danny Hillis, Freeman Dyson, Chris Langton, J. Doyne Farmer, Geoffrey West, Stuart Russell și Judea Pearl.

Un sistem emergent dinamic în desfășurare

De la întâlnirea inițială din Washington, Connecticut, până în prezent, am organizat mai multe serate și discuții în Londra și Cambridge, Massachusetts, precum și un eveniment public la Primăria Londrei. Printre invitați s-au numărat oameni de știință distinși, istorici ai științelor și teoreticieni ai comunicării, toți gândindu-se serios întreaga lor carieră la problemele inteligenței artificiale.

Am comandat eseuri unei plaje largi de participanți, cu sau fără referire la Wiener (lăsând acest lucru la latitudinea fiecărui participant). În cele din urmă, douăzeci și cinci de oameni au scris eseuri, fiecare în parte fiind îngrijorat de ce se întâmplă astăzi în era inteligenței artificiale. *Minți posibile* nu este cartea *mea*, mai degrabă este cartea *noastră*: Seth Lloyd, Judea Pearl, Stuart Russell, George Dyson, Daniel C. Dennett, Rodney Brooks, Frank Wilczek, Max Tegmark, Jaan Tallinn, Steven Pinker, David Deutsch, Tom Griffiths, Anca Dragan, Chris Anderson, David Kaiser, Neil Gershenfeld, W. Daniel Hillis, Venki Ramakrishnan, Alex „Sandy" Pentland, Hans Ulrich Obrist, Alison Gopnik, Peter Galison, George M. Church, Caroline A. Jones și Stephen Wolfram.

Văd Proiectul Minți Posibile ca pe un sistem emergent dinamic în desfășurare, o prezentare a ideilor unei comunități de gânditori sofisticați, care își pun la treabă experiența și erudiția în provocarea poveștii prevalente despre inteligența artificială digitală, comunicându-și gândurile unul altuia. Ținta este să prezentăm un

mozaic de perspective care va ajuta să clarificăm acest domeniu emergent care se dezvoltă rapid.

Le-am cerut eseiștilor să ia în considerare:

a. Poemul de tip zen „Treisprezece chipuri de a privi o mierlă"*, de Wallace Stevens, despre care insista că „nu a fost intenționat a fi o colecție de epigrame sau de idei, ci de senzații". Este un exercițiu de „perspectivism", constând în secțiuni mici și separate, fiecare menționând mierlele într-un anumit fel. Poemul este despre propria sa imaginație; privește lucrurile de care se ocupă el.
b. Parabola orbilor și elefantului. La fel ca elefantul, inteligența artificială este o temă prea mare pentru o singură perspectivă, fără a mai pune la socoteală faptul că nu există doi oameni care să privească lucrurile la fel.

Ce vrem să facă această carte? Stewart Brand a observat că „revizitarea gândirii pionierilor este întotdeauna utilă. Și oferă o perspectivă amplă, care invită la gândire decenii și chiar secole despre un subiect. Toate discuțiile contemporane sunt sortite să îmbătrânească urât și imediat fără perspectiva pe termen lung".

Danny Hillis vrea ca oamenii din domeniul inteligenței artificiale să își dea seama cum au fost programați de cartea lui Wiener. „Executați planul său", spune el, „și pur și simplu nu vă dați seama".

Dan Dennett ar dori să „îl lase pe Wiener să se ridice ca o fantomă la banchet. Gândește-te la el ca la o sursă de vigoare hibridă, o sursă de idei fără odihnă care scutură de umeri mentalitatea leneșă".

Neil Gershenfeld argumentează că „educația corectoare sub acoperire pentru oamenii care trăiesc cu trăsăturile de personalitate «Big Five» ar fi un rezultat important al cărții"

* În original, *Thirteen Ways of Looking at a Blackbird* (n.t.).

Freeman Dyson, unul dintre puținii oameni în viață care l-au cunoscut pe Wiener, observă că „*The Human Use of Human Beings* este una dintre cele mai bune cărți scrise vreodată. Wiener a nimerit aproape totul. Sunt interesat să văd ce va face cu aceste informații hoarda ta de vrăjitori".

Povestea în desfășurare a inteligenței artificiale

Lucrurile s-au schimbat și rămân la fel. Acum inteligența artificială este peste tot. Avem internetul. Avem telefoanele noastre inteligente. Fondatorii companiilor dominante – companiile care dețin „biciul ce ne biciuiește" – au venituri nete de 65 de miliarde de dolari, 90 de miliarde de dolari, 130 de miliarde de dolari. Persoane publice influente precum Elon Musk, Nick Bostrom, Martin Rees, Eliezer Yudkowsky și regretatul Stephen Hawking au emis avertismente sumbre despre inteligența artificială, care au avut ca rezultat ascensiunea unor institute bine finanțate care au sarcina de a promova „inteligența artificială bună". Dar vom putea noi, ca specie, să controlăm o inteligență artificială pe deplin realizată, nesupravegheată și care se îmbunătățește automat? Avertizările și dojanele lui Wiener din *The Human Use of Human Beings* sunt acum reale și trebuie analizate cu ochi proaspeți de cercetătorii din prima linie a revoluției inteligenței artificiale. Iată-l pe Dyson din nou:

> Wiener a respins din ce în ce mai mult ideea „veneratorilor de gadgeturi", al căror egoism corporatist a creat „motive de automatizare care trec dincolo de o curiozitate legitimă, fiind păcătoase în sine". A știut că pericolul nu este reprezentat de mașinile care se vor asemui din ce în ce mai mult cu oamenii, ci de oamenii tratați ca niște mașini. „Lumea viitorului va fi o luptă din ce în ce mai aprigă împotriva limitărilor inteligenței noastre", avertiza el în *God & Golem, Inc.*, carte publicată în 1964, în anul morții sale,

„nu un hamac confortabil, în care ne putem tolăni, în timp ce sclavii noștri roboți ne vor îndeplini toate capriciile".

Este timpul să examinăm povestea în desfășurare a inteligenței artificiale, identificându-i pe membrii principali ai comunității *mainstream*, dar și pe disidenți și prezentându-le contrapoveștile în propriile lor cuvinte.

Eseurile care urmează constituie astfel o actualizare care era necesară de mult timp în domeniu.

— John Brockman
New York, 2019

Capitolul 1
Greșit, dar mai relevant decât oricând

Seth Lloyd

Seth Lloyd este fizician teoretician la MIT, profesor Nam P. Suh în Departamentul de Inginerie Mecanică și profesor colaborator la Institutul din Santa Fe.

L-am cunoscut pe Seth Lloyd la sfârșitul anilor 1980, când moduri noi de gândire erau peste tot: importanța principiilor biologice de organizare, perspectiva informatică asupra matematicii și proceselor fizice, accentul pus pe rețelele construite în paralel, importanța dinamicii neliniare, noua înțelegere a haosului, ideile conecționiste, rețele neuronale și prelucrarea distributivă paralelă. Progresele făcute în domeniul tehnicii de calcul în acea perioadă ne-au oferit o nouă perspectivă asupra cunoașterii.

Lui Seth îi place să se numească „mecanic cuantic". Este cunoscut internațional pentru activitatea sa în domeniul tehnicii de calcul cuantice, care încearcă să controleze proprietățile exotice ale teoriei cuantice, cum ar fi superpoziția și entanglementul, pentru a rezolva probleme care ar necesita câteva vieți de om dacă ar fi rezolvate de computerele clasice.

În eseul care urmează, urmărește istoria informației de la observațiile profetice ale lui Norbert Wiener până la predicțiile unei „singularități" tehnologice, despre care unii ar vrea să credem că va înlocui specia umană. Observațiile sale privind metoda recentă de programare cunoscută sub numele de învățare profundă propun să avem un set mai modest de așteptări; spune că, în ciuda progreselor uriașe pe care le face inteligența artificială, „roboții încă nu se pot lega la șireturi".

Este dificil pentru mine să vorbesc despre Seth fără a face referire la relația sa cu prietenul și profesorul său, regretatul fizician teoretician Heinz Pagels de la Universitatea Rockefeller. Studentul și profesorul au avut, fiecare, un efect profund asupra ideilor celuilalt.

În vara anului 1988, i-am vizitat pe Heinz și Seth la Centrul de Fizică din Aspen. Activitatea lor comună pe tema complexității a fost prezentată în numărul curent din *Scientific American*; au fost exuberanți. Asta se întâmpla cu numai două săptămâni înainte de moartea tragică a lui Heinz într-un accident montan, în timp ce cobora de pe Pyramid Peak cu Seth. Vorbeau despre tehnica de calcul cuantică.

The Human Use of Human Beings, popularizarea de către Norbert Wiener din anii 1950 a influentei sale cărți *Cibernetica: sau controlul și comunicarea la animale și mașini* (1948)*, investighează interdependența dintre oameni și mașini într-o lume în care mașinile devin din ce în ce mai performante din punctul de vedere al puterii de calcul. Este o remarcabilă carte vizionară... și remarcabil de greșită. Scrisă în toiul Războiului Rece, conține o avertizare înfiorătoare despre pericolele organizațiilor și societăților totalitare, dar și ale democrației atunci când încearcă să combată totalitarismul cu arme totalitare.

Cibernetica lui Wiener analiza cu minuțiozitate științifică procesul de comandă prin *feedback* („cibernetica", din termenul grecesc antic pentru „timonier", este baza etimologică pentru cuvântul nostru** „governor", termenul pe care l-a folosit James Watt pentru a-și denumi dispozitivul revoluționar de comandă automată cu feedback care a transformat folosirea motoarelor cu abur.) Deoarece se ocupa de probleme legate de comandă, Wiener vedea lumea ca pe un set complex de cicluri *feedback* interdependente, în care senzori, semnale și dispozitive de acționare interacționează prin intermediul unui schimb complex de semnale și informații. Aplicațiile inginerești ale *ciberneticii* au avut o

* În original, *Cybernetics: or Control and Communication in the Animal and the Machine* (n.t.).

** i.e. american (n.t.).

influență și o eficiență colosale, dând naștere rachetelor, roboților, liniilor automate de asamblare și unei serii întregi de tehnici inginerești de precizie – în alte cuvinte, au pus bazele societății industriale contemporane.

Însă Wiener a avut ambiții mai mari pentru conceptele cibernetice și, în *The Human Use of Human Beings*, își prezintă ideile cu privire la aplicarea conceptelor sale în domenii diverse precum Demonul lui Maxwell*, limbajul omenesc, creier, metabolismul insectelor, sistemul legal, rolul inovației tehnologice în procesul de guvernare și chiar în domeniul religiei. Aceste aplicații mai largi ale ciberneticii au fost o nereușită aproape fără echivoc. Promovată viguros de la sfârșitul anilor 1940 până la începutul anilor 1960 – similar cu promovarea tehnologiilor informației și comunicațiilor care a condus la prăbușirea dot-com din anii 2000–2001 –, cibernetica oferea sateliți și sisteme de comutatoare pentru telefonie, însă genera puține dezvoltări utile sau chiar deloc în organizarea socială sau a societății în ansamblu.

Însă, după șaptezeci de ani, *The Human Use of Human Beings* ne poate învăța mai multe pe noi oamenii decât o putea face la început. Poate că cea mai remarcabilă caracteristică a cărții este că prezintă un număr mare de subiecte cu privire la interacțiunea dintre oameni și mașini, care au încă o relevanță considerabilă. Pe un ton sumbru, cartea face mai multe predicții despre dezastrele care urmau să apară în a doua jumătate a secolului XX, multe dintre ele fiind aproape identice cu predicțiile făcute în zilele noastre despre a doua jumătate a secolului al XXI-lea.

De exemplu, Wiener a prevăzut un moment în viitorul apropiat de anul 1950 în care oamenii ar urma să predea controlul societății unei inteligențe artificiale cibernetice, care va provoca apoi numai dezastre umanității. Automatizarea producției, preconiza Wiener, urma să creeze avantaje mari în ceea ce privește

* Demonul lui Maxwell este o ființă imaginară, inteligentă, de dimensiuni moleculare, care își permite să încalce principiul al doilea al termodinamicii (n.t.).

productivitatea, dar și să lase mulți oameni fără locuri de muncă – o secvență de evenimente care a avut într-adevăr loc în deceniile care au urmat. Dacă societatea nu urma să găsească ocupații productive pentru acești muncitori rămași fără locuri de muncă – avertiza Wiener –, vor apărea inevitabil revolte.

Dar Wiener nu a prevăzut dezvoltări tehnologice cruciale. Ca aproape toți tehnologii anilor 1950, nu a reușit să prevadă revoluția computerelor. Computerele, credea el, se vor ieftini de la sute de mii de dolari (cât costau în anii 1950) la zeci de mii de dolari; nici el, nici colegii săi de breaslă nu au anticipat explozia uriașă a puterii computerelor care va urma dezvoltării tranzistorului și circuitelor integrate. În cele din urmă, din cauza accentului pe care-l punea pe control, Wiener nu putea să conceapă o lume tehnologică în care inovația și autoorganizarea mai degrabă se ridică de la fund ca bulele în pahar decât să fie impuse de deasupra.

Concentrându-se pe răul totalitarismului (politic, științific și religios), Wiener vedea lumea într-o lumină profund pesimistă. Cartea sa ne avertiza de catastrofa care ne paște dacă nu ne corectăm comportamentul, rapid. Lumea contemporană compusă din ființe umane și mașini, la peste jumătate de secol după publicarea cărții sale, este mult mai complexă și bogată, conținând o varietate mult mai mare de sisteme politice, sociale și științifice decât a putut el să conceapă. Însă avertizările pe care ni le-a prezentat cu privire la ce se poate întâmpla dacă greșim – de exemplu, controlul întregului internet de un regim totalitar global – rămân la fel de relevante și apăsătoare astăzi pe cât erau în anii 1950.

Ce a intuit corect Wiener

Cele mai cunoscute lucrări matematice ale lui Wiener se concentrau pe probleme de analiză a semnalului și efectele zgomotului. În timpul celui de-al Doilea Război Mondial, a dezvoltat tehnici de țintire pentru avioanele de luptă, concepând modele ce puteau prognoza traiectoria viitoare a unui avion extrapolând

comportamentul său anterior. În *Cibernetica* și în *The Human Use of Human Beings*, Wiener observă că acest comportament trecut include elementele normale și anormale ale comportamentului pilotului uman, prin urmare un dispozitiv mecanizat putea prognoza comportamentul oamenilor. Ca Alan Turing, al cărui test Turing sugera că mașinile de calcul pot oferi răspunsuri la întrebări ce nu pot fi distinse de răspunsurile omenești, Wiener era fascinat de noțiunea capturării comportamentului uman în descrieri matematice. În anii 1940, și-a aplicat cunoștințele sale despre ciclurile de control și *feedback* pe *feedback*-ul neuromuscular din sistemele vii și a fost responsabil pentru aducerea lui Warren McCulloch și Walter Pitts la MIT, unde au început activitatea lor de pionierat în studiul rețelelor neuronale artificiale.

Observația principală a lui Wiener a fost că lumea trebuie înțeleasă ca informație. Sistemele complexe, cum sunt, de exemplu, organismele, creierele și societățile umane, constau în cicluri de *feedback* interconectate, în care semnalele schimbate între subsisteme rezultă în comportamente complexe, dar stabile. Când ciclurile de *feedback* nu mai funcționează, sistemul devine instabil. A construit o imagine convingătoare a modului de funcționare a funcției sistemelor biologice complexe, o imagine care este acceptată aproape universal și astăzi.

Viziunea lui Wiener asupra informației drept cantitate centrală în guvernarea comportamentului sistemelor complexe a fost remarcabilă la momentul respectiv. În zilele noastre, când automobilele și frigiderele sunt înțesate de microprocesoare și o mare parte din societatea omenească se învârte în jurul computerelor și celularelor conectate prin internet, pare prozaic să insistăm pe caracterul central al informației, tehnicii de calcul și comunicării. Pe vremea lui Wiener însă primele computere digitale abia fuseseră inventate, iar internetul nu era nici măcar o sclipire în colțul ochiului unui tehnolog.

Concepția puternică a lui Wiener conform căreia nu doar sistemele complexe proiectate, ci toate sistemele complexe au la bază cicluri de semnale și calcule a condus la contribuții majore

la dezvoltarea sistemelor complexe făcute de om. Metodele pe care el și alții le-au dezvoltat pentru controlul rachetelor, de exemplu, au fost folosite mai târziu la construirea rachetei selenare Saturn V, una dintre cele mai mari realizări inginerești din secolul XX. În special, aplicarea conceptelor cibernetice ale lui Wiener în domeniile creierului și percepției computerizate sunt precursorii direcți ai circuitelor de învățare profundă bazate pe rețele neuronale și ai inteligenței artificiale înseși. Dar dezvoltările curente din aceste domenii au deviat de la viziunea sa, iar dezvoltarea lor viitoare poate afecta temeinic atât folosirea omului de către om, cât și a mașinilor de către om.

Ce a intuit greșit Wiener

Exact în ceea ce privește extinderea ideii cibernetice la ființele umane concepțiile lui Wiener și-au ratat ținta. Lăsând la o parte deocamdată cugetările sale privind limbajul, legea și societatea omenească, priviți inovațiile mai umile, dar cu potențial utilitar, pe care le considera iminente în 1950. Wiener observă că protezele ar fi mult mai eficiente dacă purtătorii lor ar putea comunica direct cu proteza prin intermediul propriilor semnale neuronale, primind informații despre presiune și poziție de la membru și direcționând mișcarea sa ulterioară. Aceasta s-a dovedit a fi o problemă mult mai dificilă decât părea pentru Wiener: după șaptezeci de ani, protezele care să încorporeze *feedback* neuronal sunt încă în stadii incipiente de dezvoltare. Conceptul lui Wiener a fost unul extraordinar, însă este dificil să creezi o interfață între semnalele neuronale și dispozitivele electromecanice.

Mai important de atât, Wiener (alături de aproape toată lumea în anii 1950) a subapreciat masiv potențialul tehnicii de calcul digitale. Așa cum s-a mai observat, contribuțiile matematice ale lui Wiener au fost aduse pe palierul analizei semnalelor și zgomotului, iar metodele sale analitice se aplică semnalelor cu variație continuă sau analoge. Deși a participat la dezvoltarea tehnicii de calcul digitale în timpul războiului, nu a prevăzut

niciodată explozia exponențială a puterii de calcul determinată de introducerea și miniaturizarea progresivă a circuitelor cu semiconductori. Însă nu este deloc vina lui Wiener: tranzistorul nu fusese inventat încă, și tehnologia tuburilor vidate a computerelor digitale pe care o cunoștea el era grosolană, nefiabilă și nu putea fi transferată la dispozitive din ce în ce mai mari. În anexa ediției din 1948 a *Cibernetici*, anticipează computerele care vor juca șah și preconizează că vor fi capabile să gândească cu două sau chiar trei mutări înainte. Poate ar fi fost surprins să afle că, după jumătate de secol, un computer]l poate învinge pe campionul mondial uman la șah.

Supraestimarea tehnologiei și riscurile existențiale ale singularității

Când Wiener și-a scris cartea, un exemplu semnificativ de supraestimare a tehnologiei era pe cale să se producă. Anii 1950 au văzut primele eforturi în dezvoltarea inteligenței artificiale, din partea unor cercetători precum Herbert Simon, John McCarthy și Marvin Minsky, care au început să programeze computere să realizeze sarcini simple și au demarat construcția de roboți rudimentari. Succesul acestor eforturi inițiale l-a inspirat pe Simon să declare că „mașinile vor fi capabile, în decurs de douăzeci de ani, să facă orice activitate pe care o poate face și omul". Astfel de predicții s-au dovedit a fi spectaculos de greșite. Pe măsură ce au devenit mai puternice, computerele au devenit din ce în ce mai bune la jucat șah, pentru că puteau genera și evalua sistematic o selecție vastă de mutări posibile viitoare. Dar majoritatea predicțiilor despre inteligența artificială, de exemplu menajere robotice, s-au dovedit a fi iluzorii. Când Deep Blue l-a învins pe Garry Kasparov la șah în 1997, cel mai puternic robot de curățenie era Roomba, care se mișca și aspira la întâmplare, scârțâind când rămânea blocat sub canapea.

Previziunile tehnologice sunt destul de periculoase, dat fiind faptul că tehnologiile progresează printr-o serie de îmbunătățiri,

încetinite de obstacole și depășite de inovație. Multe obstacole și câteva inovații pot fi anticipate, dar și mai multe nu pot. În activitatea mea cu experimentaliști care încearcă să construiască computere cuantice observ de obicei că anumite etape tehnologice de la care mă aștept să fie ușoare se dovedesc a fi imposibile, iar alte sarcini pe care mi le imaginez a fi imposibile se dovedesc a fi ușoare. Nu știi până nu încerci.

În anii 1950, parțial inspirat de conversații cu Wiener, John von Neumann a introdus noțiunea de „singularitate tehnologică". Tehnologiile au tendința să se îmbunătățească exponențial, dublându-și puterea sau sensibilitatea la un anumit interval de timp. (De exemplu, din 1950, tehnologiile computerelor și-au dublat puterea la aproximativ doi ani, observație cunoscută ca Legea lui Moore.) Von Neumann a extrapolat rata exponențială observată a îmbunătățirilor tehnologice pentru a prezice că „progresul tehnologic va deveni incomprehensibil de rapid și complicat", depășind capacitățile omenești în viitorul nu prea îndepărtat. Într-adevăr, dacă extrapolăm creșterea puterii brute de calcul – exprimată în termeni de biți și bit-flips* – în viitor ca rata sa curentă, computerele ar trebui să ajungă din urmă creierul omenesc în următoarele două până la patru decenii (în funcție de modul de estimare a puterii de procesare a informațiilor în creierele umane).

Eșecul predicțiilor inițiale mult prea optimiste despre inteligența artificială a domolit discuțiile cu privire la singularitatea tehnologică timp de câteva decenii, dar, de la publicarea în anul 2005 a lucrării lui Ray Kurzweil *Singularitatea este aproape***, ideea progresului tehnologic care conduce la superinteligență a revenit în forță. Unii adepți ai ideii, printre care și Kurzweil, privesc singularitatea drept o oportunitate: oamenii își pot contopi creierele cu superinteligența și astfel pot trăi la nesfârșit. Alții, de exemplu Stephen Hawking și Elon Musk, au exprimat îngrijorare că această superinteligență s-ar putea dovedi malignă, privind-o

* Trecerea unui bit de la 0 la 1 sau de la 1 la 0 (n.t.).

** În original, *The Singularity IS Near* (n.t.).

ca pe cea mai mare amenințare existentă la adresa civilizației umane. Totuși, alții, printre care se numără și câțiva participanți la acest volum, consideră astfel de discuții exagerate.

Munca de o viață a lui Wiener și faptul că nu a prevăzut consecințele acesteia sunt legate intim de ideea de singularitate tehnologică iminentă. Activitatea sa în domeniul neuroștiinței și sprijinul inițial pe care l-a acordat lui McCulloch și Pitts au umbrit metodele de învățare profundă înspăimântător de eficiente din ziua de azi. În ultimul deceniu și în special în ultimii cinci ani, aceste tehnici de învățare profundă au manifestat în sfârșit ceea ce Wiener numea *Gestalt* – de exemplu, capacitatea de a recunoaște că un cerc este un cerc chiar și atunci când este aplecat într-o parte și arată ca o elipsă. Activitatea sa în domeniul comenzilor combinată cu activitatea din domeniul *feedback*-ului neuromuscular a fost semnificativă pentru dezvoltarea roboticii și este inspirația pentru interfețele dintre om și mașină bazate pe neuroni. Scăpările sale în ceea ce privește predicțiile tehnologice însă sugerează că ar trebui să fim mai atenți când discutăm despre noțiunea singularității tehnologice. Dificultățile generale ale predicțiilor tehnologice și problemele specifice dezvoltării unei superinteligențe ar trebui să ne avertizeze cu privire la supraestimarea puterii și eficienței prelucrării informațiilor.

Argumentele pentru scepticism în ceea ce privește singularitatea

Nicio creștere exponențială nu durează la nesfârșit. O explozie atomică crește exponențial, dar numai cât timp are combustibil. Similar, avansurile exponențiale din Legea lui Moore încep să atingă niște limite impuse de fizica elementară. Frecvența procesoarelor computerelor a ajuns la un maximum de câțiva gigahertzi cu un deceniu și jumătate în urmă, pur și simplu pentru că cipurile au început să se topească. Miniaturizarea tranzistorilor a întâlnit deja obstacole cuantice și mecanice din cauza fenomenului de *tunneling* și a curenților de scurgere. În cele din urmă, multiplele

îmbunătățiri exponențiale în ceea ce privește memoria și procesarea guvernate de Legea lui Moore se vor opri. Însă, probabil, în câteva decenii, puterea brută de procesare a informațiilor de care dispun computerele o va ajunge din urmă pe cea a creierului – cel puțin în măsura grosieră a numărului de biți și de *bit-flips* pe secundă.

Creierele omenești sunt construite complex, prin procesul de selecție naturală care a durat milioane de ani. Pe vremea lui Wiener, înțelegerea noastră în ceea ce privește arhitectura creierului era rudimentară și simplistă. De atunci, instrumentația și tehnicile imagistice din ce în ce mai sensibile au arătat că avem creiere cu o structură mult mai variată și o funcție mult mai complexă decât și-a putut imagina Wiener. L-am întrebat recent pe Tomaso Poggio, unul dintre pionierii neuroștiinței moderne, dacă se îngrijorează de faptul că funcția creierului uman va fi emulată curând de computere, dată fiind puterea lor crescândă de procesare a informației. „Nicio șansă", mi-a răspuns.

Progresele recente ale învățării profunde și ale tehnicii de calcul neuromorfe sunt foarte bune la reproducerea unui anumit aspect al inteligenței omenești concentrat pe funcționarea cortexului, acolo unde tiparele sunt procesate și recunoscute. Aceste progrese au permis unui computer să învingă nu numai campionul mondial la șah, cât și la Go, o realizare impresionantă, dar mai e drum lung până la robotul computerizat care să-ți facă curat în cameră. (De fapt, roboții care să se apropie de vasta capacitate umană a gamei de mișcări flexibile sunt încă un vis îndepărtat – căutați „roboți căzând"* pe internet. Roboții sunt performanți când trebuie să facă suduri de precizie pe liniile de asamblare, dar încă nu se pot lega la șireturi.)

Puterea brută de procesare a informațiilor nu presupune o putere sofisticată de procesare a informațiilor. Deși puterea computerelor a avansat exponențial, programele cu care funcționează computerele adesea nu au avansat deloc. Unul dintre principalele răspunsuri ale companiilor de software la puterea de procesare

* În original, „robots falling down" (n.t.).

sporită este să adauge caracteristici „utile", care fac adesea softul mai greu de utilizat. Microsoft Word și-a atins punctul culminant în 1995 și, de atunci, se scufundă încet sub greutatea caracteristicilor adăugate. Când Legea lui Moore va începe să încetinească, dezvoltatorii de software vor fi nevoiți să facă alegeri dificile între eficiență, viteză și funcționalitate.

O temere majoră a singularitariștilor* este că, pe măsură ce computerele se vor implica din ce în ce mai mult în proiectarea propriilor softuri, își vor încărca rapid instrucțiuni inițiale care le vor permite să atingă abilități de calcul supraomenești. Însă dovezile pe care ni le pune la dispoziție învățarea automată arată altfel. Pe măsură ce mașinile devin mai puternice și capabile de învățare, învață din ce în ce mai mult ca oamenii – din exemple multiple, adesea sub supravegherea instructorilor umani sau electronici. Educația este la fel de dificilă și lentă pentru computere cum este și pentru adolescenți. În consecință, sistemele bazate pe învățare profundă devin mai omenești și nu mai puțin omenești. Competențele pe care le aduc învățării nu sunt „mai bune decât", ci „complementare" învățării umane: sistemele computerizate de învățare pot identifica tipare pe care oamenii nu le pot identifica – și invers. Cei mai buni jucători de șah din lume nu sunt nici computere și nici oameni, ci oameni care lucrează împreună cu computerele. Cyberspațiul este într-adevăr populat de programe dăunătoare, dar acestea iau preeminent forma de *malware* – viruși notabili pentru lipsa lor malignă de rațiune, și nu pentru superinteligența lor.

Încotro, Wiener?

Wiener a observat că progresul tehnologic exponențial este un fenomen relativ modern și nu este pe de-a-ntregul bun. Privea armele atomice și dezvoltarea rachetelor cu focoase nucleare ca pe o rețetă pentru sinuciderea rasei umane. A comparat exploatarea

* În original, „singulariteers" (n.t.).

impetuoasă a resurselor planetei cu petrecerea cu ceai a Pălărierului Nebun din *Alice în Țara Minunilor*. Pustiind mediul local dintr-un anumit loc, progresăm pur și simplu mutându-ne în alt loc și pustiindu-l și pe acela. Optimismul lui Wiener cu privire la dezvoltarea computerelor și sistemelor neuromecanice a fost temperat de pesimismul său cu privire la exploatarea acestora de către guvernele asupritoare, cum ar fi cazul Uniunii Sovietice, și la tendința democrațiilor, cum ar fi cea din Statele Unite ale Americii, să devină și ele mai autoritare când se confruntă cu amenințarea autoritarismului.

Ce ar crede Wiener despre folosirea omului de către om din prezent? Ar fi uluit de puterea computerelor și a internetului. S-ar bucura că rețelele neuronale incipiente, în care a avut un rol important, au dat naștere unor sisteme puternice de învățare profundă care manifestă capacitatea perceptivă pe care o aștepta de la ele, deși, probabil, nu ar fi impresionat de faptul că unul dintre cele mai remarcabile exemple ale acestui *Gestalt* computerizat este capacitatea de a recunoaște fotografii cu pisici pe World Wide Web. În loc să privească inteligența mașinilor ca pe o amenințare, cred că ar privi-o ca pe un fenomen autonom, diferit de și coevoluând cu propria noastră inteligență umană.

Nesurprins de încălzirea globală – petrecerea cu ceai a Pălărierului Nebun din era noastră –, Wiener ar aplauda îmbunătățirea exponențială a tehnologiilor de energie alternativă, aplicându-și cunoștințele cibernetice în dezvoltarea setului complex de cicluri de *feedback* necesar pentru încorporarea acestor tehnologii în rețeaua electrică inteligentă ce urmează să apară. Cu toate acestea, recunoscând că soluția problemei încălzirii globale este cel puțin în aceeași măsură politică, pe cât este tehnologică, fără îndoială că ar fi pesimist gândindu-se la șansele noastre de a rezolva această problemă care ne amenință civilizația în timp. Wiener detesta afaceriștii – în special pe cei din lumea politică –, dar recunoștea că aceștia vor fi întotdeauna cu noi.

Este ușor să uităm cât de înspăimântătoare era lumea lui Wiener. Statele Unite ale Americii și Uniunea Sovietică erau în

plină cursă a înarmărilor, construind bombe cu hidrogen montate pe focoase nucleare transportate de rachete balistice intercontinentale ghidate de sisteme de navigare la care contribuise Wiener însuși – spre disperarea sa. Aveam patru ani când a murit Wiener. În 1964, grădinița mea se pregătea pentru un atac nuclear punându-ne să ne ascundem sub bănci. Dată fiind folosirea omului de către om din vremea sa, dacă ar putea vedea starea noastră curentă, primul răspuns al lui Wiener ar fi să răsufle ușurat că suntem încă în viață.

Capitolul 2
Limitările mașinilor de învățare opace

Judea Pearl

Judea Pearl este profesor de informatică și administrator al Laboratorului de Sisteme Cognitive de la UCLA. Cea mai recentă carte a sa, scrisă împreună cu Dana Mackenzie, este *The Book of Why: The New Science of Cause and Effect.*

În anii 1980, Judea Pearl a introdus o abordare nouă a inteligenței artificiale denumită rețelele bayesiene. Acest model bazat pe probabilități de raționare automată a permis mașinilor să funcționeze – într-o lume complexă și nesigură – ca „motoare de dovezi", revizuindu-și continuu concluziile în lumina noilor dovezi.

În decurs de câțiva ani, rețelele bayesiene ale lui Judea au întrecut complet abordările anterioare ale inteligenței artificiale bazate pe reguli. Apariția învățării profunde – în care computerele, practic, se învață singure să fie mai inteligente prin observarea multor date – l-a lăsat puțin sceptic, pentru că această metodă nu este transparentă.

Recunoscând realizările impresionante în domeniul învățării profunde din partea unor colegi precum Michael

I. Jordan și Geoffrey Hinton, se simte inconfortabil cu acest tip de opacitate. Și-a propus să înțeleagă limitările teoretice ale sistemelor de învățare profundă și evidențiază faptul că există bariere elementare care-i vor împiedica să obțină un tip de inteligență omenească, indiferent ce vor face. Folosind beneficiile de calcul ale rețelelor bayesiene, Judea și-a dat seama că o combinație de simple modele grafice și date poate fi utilizată și pentru a reprezenta și deduce relațiile cauză–efect. Importanța descoperirii sale trece cu mult dincolo de rădăcinile sale în inteligența artificială. Ultima sa carte explică gândirea cauzală pentru publicul larg; am putea spune că este un prim nivel despre cum gândim, chiar și pentru oameni.

Perspectiva principială, matematică a lui Judea asupra cauzalității, este o contribuție profundă pentru domeniul ideilor. A adus deja beneficii aproape fiecărui domeniu al cercetării, în special în științele sociale și medicale, care se bazează pe un volum mare de date.

În calitate de fost fizician, eram foarte interesat de cibernetică. Deși nu utiliza întreaga putere a mașinilor Turing, era foarte transparentă, probabil deoarece era fondată pe teoria clasică a controlului și teoria informației. Pierdem acum din această transparență, odată cu stilul de învățare profundă al învățării automate. Fundamental, este un exercițiu de ajustare a curbei ce reglează ponderile nivelurilor intermediare dintr-un lanț lung de intrare–ieșire.

Întâlnesc mulți utilizatori care spun că „funcționează bine și nu știm de ce". Când este dezlănțuită asupra seturilor mari de date, învățarea profundă are propria sa dinamică, își face propriile reparații și optimizări și îți oferă rezultatele corecte de cele mai multe ori. Dar, atunci când nu reușește să facă asta, nu ai nicio idee despre ce nu a mers și cum poți să rezolvi problema. Îndeosebi nu știi dacă eroarea este în program, în metodă sau cauzată de schimbarea stării mediului. Ar trebui să urmărim un alt fel de transparență.

Unii argumentează că transparența nu este neapărat necesară. Nu înțelegem arhitectura neuronală a creierului uman, însă funcționează bine, așa că uităm de înțelegerea noastră superficială și folosim ajutoarele umane pentru avantaje uriașe. În anumite feluri, argumentează ei, de ce să nu dezlănțuim sisteme de învățare profundă și să creăm inteligență fără să înțelegem cum funcționează? Sunt de acord cu acest argument într-o anumită măsură. Personal, nu-mi place opacitatea, așa că nu-mi voi irosi timpul cu

învățarea profundă, dar știu că-și are locul în compoziția inteligenței. Știu că sistemele netransparente pot face minuni și creierul nostru este dovada vie a acestor minuni.

Însă acest argument are limitările sale. Motivul pentru care putem să iertăm înțelegerea noastră superficială în ceea ce privește funcționarea creierului uman este că creierele noastre funcționează la fel, și acest lucru ne permite să comunicăm cu alți oameni, să învățăm de la ei, să-i instruim și să-i motivăm în limba noastră maternă. Dacă roboții noștri vor fi toți la fel de opaci ca AlphaGo, nu vom putea susține conversații profunde cu ei și ar fi trist. Ar trebui să-i reinstruim de câte ori am face o mică schimbare sarcinii sau mediului de funcționare.

Așadar, decât să experimentez cu mașinile de învățare opace, încerc să înțeleg limitările lor teoretice și să examinez cum pot fi depășite aceste limitări. Fac acest lucru în contextul sarcinilor de rațiune cauzală, care guvernează o mare parte din perspectiva oamenilor de știință asupra lumii, și, în același timp, sunt bogate în intuiție și în exemple, pentru a putea monitoriza progresul analizei noastre. În acest context, am descoperit că există câteva bariere elementare și, dacă nu le depășim cumva, nu ne vom apropia măcar de un soi de inteligență umană, indiferent ce facem. După mine, cartografierea acestor bariere nu este mai puțin importantă decât să ne lovim de ele.

Sistemele curente de învățare automată funcționează aproape exclusiv într-un mod statistic sau cu model orb, lucru analog în multe feluri încercării de a implementa o funcție într-un *cloud* de puncte de date. Astfel de sisteme nu își pot pune întrebarea „Ce-ar fi dacă?", neputând servi așadar drept bază pentru o inteligență artificială puternică – și anume, o inteligență artificială care emulează rațiunea și competența umane. Pentru a ajunge la nivelul inteligenței umane, mașinile de învățare au nevoie de ghidarea unei scheme a realității, unui model – similar unei hărți care ne ghidează atunci când conducem printr-un oraș pe care nu-l cunoaștem.

Pentru a fi mai clar, mașinile curente de învățare își îmbunătățesc performanța optimizând parametri pentru un flux de intrări senzoriale primite de la mediu. Este un proces lent, analog procesului de selecție naturală care stă la baza evoluției darwiniste. Explică, astfel, cum specii precum vulturii și șerpii au dezvoltat sisteme de văz excelente de-a lungul a milioane de ani, însă nu poate explica procesul supraevoluționar care permite oamenilor să producă ochelari și telescoape în decurs de mai puțin de o mie de ani. Ce au avut oamenii și nu au avut alte specii a fost o reprezentare mentală a mediului lor – o reprezentare pe care au putut să o manipuleze după propria voință pentru a-și imagina medii ipotetice alternative pentru planificare și învățare.

Istorici ai lui homo sapiens, cum ar fi Yuval Noah Harari și Steven Mithen, sunt de acord că ingredientul decisiv care a conferit strămoșilor noștri capacitatea de a obține dominația asupra întregii planete, cu aproximativ 40 000 de ani în urmă, a fost capacitatea lor de a crea și memora o reprezentare mentală a mediului înconjurător, de a interoga acea reprezentare, de a o distorsiona prin acte mentale ale imaginației și de a răspunde în cele din urmă întrebărilor care încep cu „Dar dacă?" Exemple sunt întrebările intervenționiste („Dacă fac așa și așa?") și întrebările retrospective și contrafactuale („Dacă aș fi făcut altceva?"). Nicio mașină de învățare în funcțiune astăzi nu poate răspunde la aceste întrebări. Mai mult chiar, majoritatea mașinilor de învățare nu posedă o reprezentare din care să derive răspunsurile la astfel de întrebări.

Cu privire la rațiunea cauzală, constatăm că poți face foarte puține lucruri cu orice fel de ajustare a curbei cu model orb sau orice inferență statistică, indiferent cât este de sofisticat procesul de ajustare. De asemenea, am găsit un cadru teoretic pentru organizarea acestor limitări, care formează o ierarhie.

La primul nivel, ai gândirea statistică, ce-ți poate spune doar cum percepția unui eveniment îți poate schimba perspectiva asupra altuia. De exemplu, ce-ți poate spune un simptom despre o boală?

Apoi ai un al doilea nivel, care-l implică pe primul, dar nu și invers. Aici este vorba despre acțiuni. „Ce s-ar întâmpla dacă am crește prețurile?" „Dacă mă faci să râd?" Acest al doilea nivel al ierarhiei necesită informații despre intervențiile care nu sunt disponibile în primul. Aceste informații pot fi codate într-un model grafic, care ne spune doar ce variabilă corespunde alteia.

Al treilea nivel al ierarhiei este cel contrafactual. Acesta este limbajul utilizat de oamenii de știință. „Dacă obiectul ar fi de două ori mai greu?" „Dacă aș face lucrurile altfel?" „Durerea de cap mi-a trecut de la aspirină sau de la somn?" Enunțurile contrafactuale sunt pe ultimul nivel în sensul că nu pot fi derivate nici dacă am putea prezice efectele tuturor acțiunilor. Au nevoie de un ingredient suplimentar, sub forma ecuațiilor, pentru a ne spune cum răspund variabilele la modificările altor variabile.

Una dintre cele mai mari realizări ale cercetării inferenței cauzale a fost algoritmizarea intervențiilor și enunțurilor contrafactuale, ultimele două niveluri ale ierarhiei. Cu alte cuvinte, după ce codăm cunoștințele noastre științifice într-un model (care poate fi calitativ), există algoritmi care examinează modelul și stabilesc dacă o anumită interogație, fie ea despre intervenție sau despre enunțul contrafactual, poate fi estimată din datele disponibile – și, dacă da, cum. Această capacitate a transfigurat modul în care oamenii de știință se ocupă de știință, în special în cele în care informațiile abundă, cum sunt sociologia și epidemiologia, pentru care modelele cauzale au devenit o a doua limbă. Aceste discipline văd transformările lor lingvistice ca pe Revoluția Cauzală. Cum spune cercetătorul de științe sociale de la Harvard, Gary King, „s-au învățat mai multe despre inferența cauzală în ultimele decenii decât s-a învățat în toată istoria înregistrată până atunci".

Contemplând succesul învățării automate și încercând să-l extrapolez la viitorul inteligenței artificiale, mă întreb: suntem conștienți de limitările elementare care au fost descoperite în arena inferenței cauzale? Suntem pregătiți să facem față impedimentelor teoretice care ne împiedică să trecem de la un nivel al ierarhiei la altul?

Eu privesc învățarea automată ca pe un instrument care ne conduce de la date la probabilități. Dar apoi trebuie să mai facem încă doi pași pentru a trece de la probabilități la înțelegerea reală – doi pași mari. Unul este să preconizăm efectul acțiunilor, iar al doilea este imaginația contrafactuală. Nu putem pretinde că înțelegem realitatea decât dacă facem ultimii doi pași.

În pătrunzătoarea sa carte *Foresight and Understanding* (1961), filosoful Stephen Toulmin a identificat contrastul dintre transparență și opacitate drept cheia înțelegerii rivalității antice dintre știința greacă și babiloniană. După părerea lui Toulmin, astronomii babilonieni erau maeștri ai predicțiilor de tip cutie neagră, întrecându-i lejer pe rivalii lor greci în ceea ce privește acuratețea și concordanța observațiilor celeste. Însă știința a favorizat strategia creativ-speculativă a astronomilor greci, care era întreținută puternic cu imagistica metaforică: tuburi circulare pline de foc, orificii mici prin care focul celest era vizibil sub formă de stea și pământul emisferic purtat pe spatele unei broaște-țestoase uriașe. Această strategie de modelare excentrică, și nu extrapolarea babiloniană a fost cea care i-a dat imboldul lui Eratostene (276–194 î.Hr.) să realizeze unul dintre cele mai creative experimente din lumea antică și să calculeze circumferința Pământului. Un astfel de experiment nu i-ar fi trecut niciodată prin minte unui arhivator de date babilonian.

Abordările cu model orb impun limitări intrinseci sarcinilor cognitive pe care inteligența artificială puternică le poate îndeplini. Concluzia mea generală este că inteligența artificială de nivel uman nu poate lua naștere doar din mașini de învățare cu model orb; necesită colaborarea simbiotică a datelor și modelelor.

Știința datelor este o știință doar în măsura în care facilitează interpretarea datelor – o problemă cu două fațete, conectarea datelor la realitate. Datele în sine nu sunt o știință, indiferent cât de „mari" ajung și de iscusința cu care sunt manipulate. Sistemele opace de învățare ne pot duce până în Babilon, dar nu până în Atena.

Capitolul 3
Scopul scris în mașină

Stuart Russell

Stuart Russell este profesor de informatică și profesor de inginerie Smith-Zadeh la Universitatea Berkeley. Este coautorul (alături de Peter Norvig) al cărții *Artificial Intelligence: A Modern Approach.*

Informaticianul Stuart Russell alături de Elon Musk, Stephen Hawking, Max Tegmark și mulți alții au insistat să se acorde atenție potențialelor pericole implicate de crearea unei inteligențe de nivel supraomenesc (sau chiar omenesc) – o inteligență artificială generală, ale cărei scopuri programate nu se vor alinia neapărat cu ale noastre.

Cărțile sale anterioare au fost pe tema înțelegerii noțiunii de „caracter optim legat" ca definiție formală a inteligenței pe care se poate lucra. A dezvoltat tehnica metaraționării raționale, „care este, în termeni largi, de a face calculele așteptate pentru îmbunătățirea calității deciziei tale finale cât se poate de repede". De asemenea, a lucrat la unificarea teoriei probabilității și logicii de gradul întâi – ceea ce a condus la crearea unui sistem de monitorizare nou și mult mai eficient pentru Tratatul privind interzicerea

totală a testelor nucleare – și la problema procesului de luare a deciziilor pe perioade lungi de timp (prezentările sale pe această ultimă temă se intitulează, de obicei „Viața: joacă și câștigă în 20 de trilioane de mutări").

Este foarte îngrijorat de dezvoltarea continuă a armelor autonome, cum ar fi microdronele letale, care pot fi transformate în arme de distrugere în masă. A redactat o scrisoare de la patruzeci de cercetători de top din domeniul inteligenței artificiale și a trimis-o președintelui Obama, determinând astfel o serie de ședințe de securitate națională la nivel înalt.

Activitatea sa din prezent se concentrează pe crearea a ceea ce el numește inteligență artificială „vădit benefică". Dorește să creeze siguranța inteligenței artificiale „impregnând sistemele cu nesiguranță explicită" cu privire la obiectivele programatorilor lor umani, o abordare care ar presupune o reorganizare radicală transparentă a cercetării inteligenței artificiale din prezent.

Stuart este, de asemenea, și pe radarul oricărei persoane care a făcut un curs universitar de informatică în ultimii douăzeci și ceva de ani. Este coautorul celui mai cunoscut manual pe tema inteligenței artificiale, cu un număr estimat de peste 5 milioane de cititori în limba engleză.

Printre multele probleme semnalate în cartea lui Norbert Wiener *Folosirea omului de către om* (1950) care sunt relevante în prezent, cea mai importantă pentru cercetătorul inteligenței artificiale este posibilitatea ca umanitatea să cedeze mașinilor controlul asupra destinului său.

Wiener a considerat mașinile din viitorul apropiat mult prea limitate pentru a exercita un control global, imaginându-și, în schimb, că mașinile și sistemele de control asemănătoare mașinilor vor fi controlate de elitele omenești pentru a reduce marea masă a umanității la starea de „rotițe, manete și tije". Privind și mai mult în viitor, a evidențiat dificultatea de a specifica absolut corect obiectivele pentru mașinile cu capacități mari, observând

> „adevăruri simple și evidente ale vieții, cum ar fi djinnul găsit într-o sticlă, pe care ai face bine să-l lași acolo; al pescarului care-și dorește prea multe binefaceri de la Dumnezeu pentru soția sa și care ajunge fix acolo de unde a plecat; al faptului că, dacă ți se acordă trei dorințe, trebuie să fii foarte atent ce-ți dorești".

Pericolele sunt clare:

> „Vai de noi dacă lăsăm «mașina» să ne decidă comportamentul înainte de a examina legile acțiunilor sale și de a ști bine dacă acest comportament al mașinii va fi ghidat de

principii acceptabile pentru noi! Pe de altă parte, mașina, asemenea djinnului, care poate învăța și lua decizii în baza a ceea ce a învățat, nu va fi obligată în niciun fel să ia decizii pe care noi ar fi trebuit să le luăm sau care să fie acceptabile pentru noi".

Zece ani mai târziu, după ce a văzut cum programul de jucat table al lui Arthur Samuel învață să joace table mult mai bine decât creatorul său, Wiener a publicat articolul „Câteva consecințe morale și tehnice ale automatizării" în revista *Science*. În această lucrare, mesajul este și mai clar:

„Dacă, pentru a ne atinge scopurile, folosim un agent mecanic cu a cărui funcționare nu putem interfera în mod eficient... ar fi bine să fim foarte siguri că scopul pe care îl programăm în mașini este scopul pe care-l dorim cu adevărat".

Din punctul meu de vedere, aceasta este sursa riscului existențial al inteligențelor artificiale superinteligente citat în ultimii ani de observatori precum Elon Musk, Bill Gates, Stephen Hawking și Nick Bostrom.

Programarea scopurilor în mașini

Obiectivul cercetării inteligenței artificiale a fost înțelegerea principiilor care stau la baza comportamentului inteligent și integrarea acelor principii în mașini care pot manifesta un astfel de comportament. În anii 1960 și 1970, noțiunea prevalentă de inteligență era capacitatea de gândire logică, ce includea abilitatea de a deriva planuri de acțiune ce puteau atinge garantat un obiectiv specificat. Recent, s-a creat un consens în jurul ideii de agent rațional care percepe și acționează în conformitate pentru a-și maximiza utilitatea prevăzută. Subdomenii, cum ar fi planificarea logică, robotica și înțelegerea limbajului natural, sunt cazuri

speciale de paradigme generale. Inteligența artificială a încorporat teoria probabilității pentru a face față nesiguranței, teoria utilității pentru a defini obiectivele și învățarea statistică pentru a permite mașinilor să se adapteze la circumstanțe noi. Aceste dezvoltări au creat legături puternice cu alte discipline care se bazează pe concepte similare, printre care teoria controlului, economia, cercetarea operațiilor și statistica.

Atât din perspectiva de planificare logică, precum și din cea de agent rațional asupra inteligenței artificiale, obiectivul mașinii – fie acesta sub formă de scop, funcție utilitară sau funcție de recompensare (ca în învățarea cu întărire) – este specificat din exterior. În cuvintele lui Wiener, acesta este „scopul programat în mașină". Într-adevăr, este una dintre regulile de bază ale domeniului ca sistemele de inteligență artificială să servească scopuri generale – adică să fie capabilă să accepte un scop ca intrare și apoi să-l atingă –, și nu scopuri speciale, cu obiectivul stabilit implicit în designul lor. De exemplu, un autoturism cu pilot automat ar trebui să accepte o destinație ca intrare, și nu să aibă o destinație fixă. Însă unele aspecte ale „scopului de șofat" ale autoturismului sunt fixe, cum ar fi faptul că nu trebuie să lovească pietonii. Acesta este construit direct în algoritmii de șofat ai autoturismului, nefiind explicit: niciun autoturism cu pilot automat care există în zilele noastre nu „știe" că pietonii preferă să nu fie călcați.

Programarea unui scop într-o mașină pentru a-i optimiza comportamentul conform unor algoritmi clar definiți pare să fie o abordare admirabilă pentru a ne asigura că, așadar, „comportamentul mașinii se va desfășura conform unor principii acceptabile pentru noi!" Dar, așa cum avertizează Wiener, trebuie să introducem scopul potrivit. Am putea denumi această problemă problema regelui Midas: Midas a primit exact ce a cerut – și anume ca tot ce atinge să se transforme în aur –, dar a descoperit prea târziu dezavantajele de a bea aur lichid și de a mânca aur solid. Termenul tehnic pentru programarea scopului potrivit este alinierea valorii. Când eșuează, putem insufla accidental mașinilor obiective

contrare cu ale noastre. Dacă sistemul de inteligență artificială primește sarcina de a găsi un tratament pentru cancer cât se poate de repede, acesta ar putea alege să folosească întreaga populație umană drept cobai pentru experimentele sale. Dacă i se cere să amelioreze aciditatea din oceane, ar putea folosi tot oxigenul din atmosferă ca efect secundar. Aceasta este o caracteristică obișnuită a sistemelor care folosesc optimizarea: variabilele care nu sunt incluse în obiectiv pot fi setate la valori extreme pentru a ajuta la optimizarea obiectivului respectiv.

Din nefericire, nici inteligența artificială și nici alte discipline (economia, statistica, teoria controlului, cercetarea operațiilor) structurate în jurul optimizării obiectivelor nu au multe de spus despre metoda de identificare a scopurilor „pe care ni le dorim cu adevărat". În schimb, acestea presupun că obiectivele sunt pur și simplu implantate în mașină. Cercetarea inteligenței artificiale, în forma sa actuală, studiază abilitatea de a atinge obiective, și nu proiectarea obiectivelor respective.

Steve Omohundro a arătat încă o dificultate, observând că entitățile inteligente trebuie să acționeze pentru a-și prezerva propria existență. Această tendință nu are nimic de-a face cu instinctul de autoconservare sau altă noțiune biologică; pur și simplu, o entitate nu-și poate atinge obiectivul dacă este moartă. Conform argumentului lui Omohundro, o mașină superinteligentă care are un comutator de oprire – pe care unii, inclusiv Alan Turing într-o prelegere susținută în 1951 la BBC Radio 3, au văzut-o ca pe potențiala noastră salvare – va lua măsuri pentru a dezactiva acest comutator într-o anumită măsură.* Astfel putem face față ideii de mașini superinteligente – acțiunile lor fiind prin definiție imprevizibile pentru noi, iar obiectivele lor imperfect specificate intrând în conflict cu ale noastre –, ale căror motivații de a-și prezerva existența pentru a atinge acele obiective ar putea fi de neînvins.

---

* Omohundro, „The Basic AI Drives", in *Proceedings of the First AGI Conference*, 171; and in P. Wang, B. Goertzel, and S. Franklin, ed., *Artificial General Intelligence* (Amsterdam, The Netherlands: IOS Press, 2008).

1 001 motive de a nu fi atenți

Au fost emise obiecții la aceste argumente, în primul rând de către cercetătorii din comunitatea inteligenței artificiale. Obiecțiile reflectă o reacție defensivă naturală, cuplată probabil cu lipsa imaginației despre ce ar putea face o mașină superinteligentă. Niciuna nu pare să rămână în picioare la o examinare mai atentă. Iată câteva dintre cele mai comune:

- *Nu-ți face probleme, o putem dezactiva pur și simplu**. Acesta este adesea primul lucru care apare în mintea unui diletant când ia în considerare riscurile inteligenței artificiale superinteligente – ca și cum o entitate superinteligentă nu s-ar gândi niciodată la asta. Aceasta se aseamănă mai degrabă cu a spune că riscul de a pierde în fața lui Deep Blue sau AlphaGo este neglijabil – tot ce trebuie să facem este să facem mutările potrivite.
- *Inteligența artificială de nivel omenesc sau supraomenesc este imposibilă*** Aceasta este o afirmație neobișnuită pentru cercetătorii din domeniul inteligenței artificiale, dat fiind faptul că, de la Turing încoace, s-au apărat de asemenea acuzații venite de la filosofi și matematicieni. Afirmația, care nu are la bază nicio dovadă, pare să accepte că, dacă inteligența artificială superinteligentă *ar fi* posibilă, *ar fi* un risc semnificativ. Este ca și cum un șofer de autobuz, având ca pasageri toată umanitatea, ar spune: „Da, conduc către o

* Cercetătorul din domeniul inteligenței artificiale Jeff Hawkins, de exemplu, scrie „Anumite mașini inteligente vor fi virtuale, ceea ce înseamnă că vor exista și acționa doar în rețelele de calculatoare... Întotdeauna este posibil să oprești o rețea de computere, chiar dacă este dureros", https://www.recode.net/2015/3/2/11559576/ (n.a.).

** Raportul AI100 (Peter Stone et al.), sponsorizat de Universitatea Stanford, include următoarele: „Spre deosebire de filme, nu se întrezărește nicio rasă de roboți supraumani la orizont și, probabil, nici nu este posibil", https://ai100.stanford.edu/2016-report (n.a.).

prăpastie – de fapt, o să apăs pedala până la blană! Dar ai încredere în mine, o să rămânem fără combustibil înainte de a ajunge!" Afirmația reprezintă un pariu nesăbuit contra ingeniozității umane. Am mai făcut astfel de pariuri și am pierdut. În data de 11 septembrie 1933, renumitul fizician Ernest Rutherford a afirmat cu încredere deplină: „Oricine se așteaptă să descoperim o sursă de energie în transformarea acestor atomi vorbește prostii". În data de 12 septembrie 1933, Leo Szilard a inventat reacția nucleară în lanț indusă de neutroni. La câțiva ani, a demonstrat o astfel de reacție în laboratorul său de la Universitatea din Columbia. După cum își amintea într-un memoriu: „Am oprit totul și am mers acasă. În noaptea aceea, în mintea mea mai încăpea foarte puțină îndoială că lumea se îndrepta către suferință".

- *Este prea devreme să ne facem probleme.* Momentul potrivit să ne îngrijorăm de o potențială problemă gravă a umanității nu depinde numai de momentul în care va apărea problema, ci și de cât timp este necesar pentru a dezvolta și implementa o soluție care evită riscurile. De exemplu, dacă ar fi să detectăm un asteroid uriaș care urmează să lovească Pământul în 2067, am spune oare: „Este prea devreme să ne îngrijorăm"? Și, dacă analizăm riscurile globale catastrofale ale încălzirii globale care urmează să aibă loc la sfârșitul acestui secol, este oare prea devreme să luăm măsuri pentru a le preveni? Dimpotrivă, este posibil să fie prea târziu. Perioada de timp relevantă pentru inteligența artificială de nivel omenesc este mai puțin predictibilă, dar, ca fisiunea nucleară, poate sosi mult mai devreme decât ne așteptăm. O variație a acestui argument este declarația lu Andrew Ng că ar fi „ca și cum ne-am îngrijora de suprapopularea planetei Marte". Aceasta trimite la o analogie convenabilă: nu doar că riscul este ușor de gestionat și departe în viitor, dar este și extrem de improbabil să încercăm să mutăm miliarde de oameni pe Marte. Analogia este falsă, însă. Deja dedicăm resurse tehnice și științifice considerabile creării sistemelor

de inteligență artificială din ce în ce mai performante. O analogie mai potrivită ar fi să plănuim să mutăm rasa umană pe Marte fără a lua în considerare ce vom mânca, ce vom bea și ce vom respira după ce ajungem.

- *Inteligența artificială de nivel omenesc nu este iminentă, în orice caz.* Raportul AI100, de exemplu, ne asigură: „Contrar preconizărilor fantastice pentru inteligența artificială în presa populară, Comitetul de Studiu nu a găsit nicio cauză de îngrijorare care să arate că inteligența artificială este o amenințare iminentă pentru omenire". Acest argument pur și simplu greșește motivele de îngrijorare, care nu sunt predicate pe iminență. În cartea sa din 2014, *Superintelligence: Paths, Dangers, Strategies,* Nick Bostrom, de exemplu, scrie: „Nu face parte din argumentația acestei cărți că suntem pe cale să facem o descoperire revoluționară în inteligența artificială sau că putem prezice cu precizie când ar putea apărea o astfel de dezvoltare".
- *Ești doar un luddit.* Este o definiție ciudată a ludditului care-i include pe Turing, Wiener, Minsky, Musk și Gates, care sunt printre cei mai prolifici participanți la progresul tehnologic din secolele al XX-lea și al XXI-lea*. În plus, epitetul reprezintă o neînțelegere completă a naturii îngrijorărilor emise și a scopului pentru care au fost emise. Este ca și cum cineva ar acuza inginerii nucleari de luddism dacă ar evidenția nevoia de a controla reacția de fisiune. Unii contestatari folosesc și termenul „anti-AI", care este ca și cum ai spune despre inginerii nucleari că sunt „antifizică". Scopul înțelegerii și împiedicării riscurilor inteligenței artificiale este să ne asigurăm că ne putem bucura de beneficii. Bostrom, de exemplu, scrie că succesul controlării

* Elon Musk, Stephen Hawking și alții (printre care, se pare, și autorul) au primit în 2015 premiul Ludditul Anului de la Fundația de Inovație în Tehnologia Informației: https://itif.org/publications/2016/01/19/artificial-intelligence-alarmists-win-itif%E2%80%99s-annual-luddite-award (n.a.).

inteligenței artificiale va avea ca efect „o traiectorie civilizațională care va conduce la o folosire compasională și jubilantă a dotărilor cosmice ale umanității" – o predicție deloc pesimistă.

- *Orice mașină care este suficient de inteligentă să cauzeze probleme va fi suficient de inteligentă să aibă obiective corespunzătoare și altruiste**. (Adesea, argumentul adaugă premisa că oamenii cu inteligență superioară au tendința de a avea obiective mai altruiste, o perspectivă care poate fi legată mai mult de sinele celor care aduc argumentul.) Acest argument este legat de problema lui Hume privind discrepanța între ceea ce este și ceea ce ar trebui să fie** și eroarea naturalistă a lui G.E. Moore, care sugerează că mașina, ca urmare a inteligenței sale, va *percepe* cumva ce este bine, dată fiind experiența ei în lume. Acest lucru este neplauzibil; de exemplu, un om nu poate deduce obiectivul de șah mat privind tabla de șah sau piesele de șah; aceeași tablă de șah și aceleași piese pot fi utilizate pentru antișah sau multe alte jocuri care nu au fost inventate încă. Cu alte cuvinte: în timp ce Bostrom își imaginează oamenii duși la extincție de un robot imaginar care transformă planeta într-o mare de agrafe pentru hârtie, noi oamenii vedem acest rezultat ca tragic, în timp ce bacteria *thiobacillus ferrooxidans* care consumă fier este încântată. Cine poate spune că bacteria greșește? Faptul că mașina a primit un obiectiv clar din partea oamenilor nu înseamnă că va recunoaște automat importanța oamenilor ca lucruri care nu fac parte din obiectiv. Maximizarea obiectivului poate

* Rodney Brooks, de exemplu, spune că este imposibil ca un program să fie „suficient de inteligent astfel încât să poată inventa moduri de a submina societatea umană pentru a atinge scopuri stabilite pentru el de către oameni, fără a înțelege modurile în care cauzează probleme pentru aceiași oameni", http://rodneybrooks.com/the-seven-deadly-sins-of-predicting-the-future-of-ai (n.a.).

** În original, „is-ought problem" (n.t.).

cauza probleme oamenilor, dar, prin definiție, mașina nu va recunoaște acele probleme ca problematice.

- *Inteligența este multidimensională, deci „mai inteligentă decât a oamenilor" este un concept fără sens**. Este o idee principală a psihologiei moderne că IQ-ul nu reprezintă întregul spectru de competențe cognitive pe care le posedă oamenii în diverse grade. IQ-ul este într-adevăr o măsurătoare lipsită de precizie a inteligenței umane, dar nu are niciun sens pentru sistemele moderne de inteligență artificială, deoarece capacitățile lor din diverse domenii nu sunt corelate. Cum putem compara IQ-ul motorului de căutare Google, care nu poate juca șah, cu cel al lui Deep Blue, care nu poate răspunde cererilor de căutare?
- Niciun motiv nu susține argumentul conform căruia, dacă inteligența este multilaterală, putem ignora riscul mașinilor superinteligente. Dacă „mai inteligent decât oamenii" este un concept fără sens, atunci la fel este și „mai inteligent decât gorilele" și, prin urmare, gorilele nu au niciun motiv de îngrijorare în ceea ce-i privește pe oameni. În mod clar argumentul nu este valabil. Nu doar că este posibil pentru o entitate să fie mai capabilă decât alta în toate dimensiunile relevante ale inteligenței, este posibil și ca o specie să reprezinte o amenințare existențială pentru alta, chiar dacă celei din urmă îi lipsește aprecierea pentru muzică și literatură.

Soluții

Putem oare să abordăm avertizările lui Wiener direct? Putem proiecta sisteme de inteligență artificială ale căror scopuri nu intră în conflict cu ale noastre, pentru a putea fi mulțumiți de modul în care se comportă? La prima vedere, pare fără speranță, pentru că se va dovedi cu siguranță nerealist să ne scriem corect scopurile

* Kevin Kelly „The Myth of a Superhuman AI", *Wired*, April 25, 2017 (n.a.).

sau să ne imaginăm toate modurile contraintuitive prin care o entitate superinteligentă le poate atinge.

Dacă tratăm sistemele de inteligență artificială superinteligentă ca și cum ar fi cutii negre venite din spațiu, atunci, într-adevăr, nu mai avem nicio speranță. În schimb, abordarea pe care se pare că suntem obligați să o folosim, dacă e să avem orice fel de încredere în rezultat, este să definim o problemă formală *F* și să proiectăm sisteme de inteligență artificială care să poată rezolva probleme de tipul *F*, astfel încât, indiferent de cât de bine rezolvă sistemul problemele *F*, noi vom fi mulțumiți în mod garantat cu soluția. Dacă putem să concepem un *F* corespunzător care să aibă această proprietate, vom putea să creăm o inteligență artificială *despre care putem dovedi că este benefică.*

Iată un exemplu pentru cum *nu* ar trebui să procedăm: fie ca o recompensă de valoare scalară să fie furnizată periodic de către om mașinii, în funcție de cât de bine s-a comportat mașina în perioada respectivă, și fie *F* problema maximizării sumei de recompense așteptate de mașină. Soluția optimă la această problemă nu este, așa cum ar putea spera unii, să te comporți bine, ci să preiei controlul asupra omului și să-l forțezi să-ți ofere un flux de recompense maxime. Aceasta este cunoscută drept problema *wireheading*, bazată pe observații conform cărora oamenii înșiși sunt susceptibili să întâmpine aceeași problemă dacă centrele plăcerii din creier le sunt stimulate electronic.

Există, consider eu, o abordare care ar putea să funcționeze. Oamenii pot fi descriși în mod rezonabil ca având preferințe (de cele mai multe ori implicite) pentru viețile lor viitoare – adică, dacă i se acordă suficient timp și ajutoare vizuale nelimitate, omul poate exprima o preferință (sau indiferență) când i se oferă alegerea între două vieți viitoare care i se deschid în toate aspectele. (Această idealizare ignoră posibilitatea ca mințile noastre să fie compuse din subsisteme cu preferințe incompatibile; dacă este adevărat, acest lucru ar limita abilitatea unei mașini de a ne satisface optim preferințele, dar nu pare să ne împiedice să proiectăm mașini ce evită rezultate catastrofale.) Problema formală *F* care

trebuie rezolvată de mașină în acest caz este maximizarea preferințelor vieții viitoare a omului, neștiind inițial cu siguranță care sunt acestea. În plus, deși preferințele vieții viitoare sunt variabile ascunse, sunt înrădăcinate într-o sursă voluminoasă de dovezi – și anume, toate alegerile omenești făcute vreodată. Această formulă evită problema lui Wiener: mașina poate învăța mai multe despre preferințele umane în timp, bineînțeles, dar nu va atinge niciodată certitudinea.

O definiție mai precisă este dată de cadrul învățării cooperative cu întărire inversă* sau CIRL. O problemă CIRL implică doi agenți: unul uman și celălalt robot. Deoarece există doi agenți, problema este ceea ce economiștii numesc un joc. Este un joc de informare parțială, deoarece omul cunoaște funcția de recompensare, iar robotul nu – chiar dacă funcția robotului este să o maximizeze.

Un exemplu simplu: să presupunem că lui Harriet, omul, îi place să colecteze agrafe de hârtie și capse și că funcția ei de recompensare depinde de cât de multe are din fiecare. Mai exact, dacă are p agrafe și s capse, gradul său de fericire este $\theta p + (1-\theta)s$, unde θ este, în esență, un curs de schimb între agrafele de hârtie și capse. Dacă θ este 1, înseamnă că-i plac numai agrafele de hârtie; dacă θ este 0, îi plac doar capsele; dacă θ este 0,5, îi plac ambele în aceeași măsură și așa mai departe. Este sarcina lui Robby, robotul, să producă agrafele de hârtie și capsele. Obiectivul jocului este că Robby vrea să o facă pe Harriet fericită, dar nu cunoaște valoarea lui θ, așa că nu știe sigur cât de multe din fiecare trebuie să producă.

Iată cum funcționează jocul. Fie valoarea reală a lui θ 0,49 – adică Harriet are o mică preferință pentru capse în detrimentul agrafelor de hârtie. Și să presupunem că Robby are o bănuială anterioară uniformă despre θ – adică, crede că θ poate fi în egală măsură orice valoare între 0 și 1. Harriet face acum o mică demonstrație, producând două agrafe, două capse sau câte una din

* În original, „cooperative inverse-reinforcement learning" (n.t.).

fiecare. Apoi, robotul poate produce nouăzeci de agrafe, nouăzeci de capse sau cincizeci din fiecare. Ai putea crede că Harriet, care preferă capsele, ar trebui să producă două capse. Dar, în acest caz, răspunsul natural al lui Robby ar fi să producă nouăzeci de capse (cu o valoare totală pentru Harriet de 49,5), care este un rezultat mai puțin dorit pentru Harriet decât cincizeci din fiecare (valoare totală 50,0). Soluția optimă a acestui joc este ca Harriet să producă câte una din fiecare, pentru ca astfel Robby să producă câte cincizeci din fiecare. Astfel, modul în care este definit jocul o încurajează pe Harriet să-l „învețe" pe Robby – atât timp cât știe că Robby o urmărește cu atenție.

În cadrul CIRL, se poate formula și rezolva problema comutatorului de oprire – adică, problema metodei prin care poți împiedica robotul să-și dezactiveze comutatorul de oprire. (Turing poate răsufla liniștit.) Un robot care este nesigur de preferințele omului beneficiază de pe urma faptului că este oprit, pentru că înțelege că omul va apăsa pe comutatorul de oprire pentru a împiedica robotul să facă ceva contrar acelor preferințe. Astfel robotul este stimulat să păstreze comutatorul de oprire și stimulentul său derivă direct din nesiguranța sa cu privire la preferințele umane.*

Exemplul comutatorului de oprire sugerează șabloane pentru designuri cu agent controlabil și pune la dispoziție cel puțin un caz de sistem despre care se poate dovedi că este benefic în sensul prezentat mai sus. Abordarea generală seamănă cu problemele de proiectare a mecanismelor din economie, în care o persoană stimulează alți agenți să se comporte în moduri benefice pentru proiectant. Diferența-cheie de aici este că noi construim unul dintre agenți pentru a servi celuilalt.

Există motive să credem că această abordare poate funcționa practic. În primul rând există informații abundente scrise și filmate despre oameni făcând lucruri (și despre alți oameni care reacționează). Tehnologia de construire a modelelor de preferințe

* Consultați Hadfield-Menell et al., „The Off- Switch Game", https://arxiv.org/pdf/1611.08219.pdf (n.a.).

umane din acest depozit va fi probabil disponibilă cu mult timp înainte de crearea sistemelor de inteligență artificială superinteligentă. În al doilea rând, există stimulente economice pe termen scurt pentru ca roboții să înțeleagă preferințele umane: dacă un robot domestic proiectat prost gătește pisica la cină, fără să-și dea seama că valoarea ei sentimentală este mai mare decât valoarea ei nutrițională, industria roboților domestici va da faliment.

Însă există dificultăți clare cu o abordare care se așteaptă de la robot să învețe preferințele inerente din comportamentul omului. Oamenii sunt irațional, inconsecvenți, au voința slabă și sunt limitați din punctul de vedere al puterii de calcul, așa că acțiunile lor nu reflectă întotdeauna adevăratele lor preferințe. (Gândește-te, de exemplu, la doi oameni care joacă șah. De obicei, unul dintre ei pierde, dar nu dinadins!) Așadar roboții pot învăța din comportamentul uman nerațional doar cu ajutorul unor modele cognitive ale oamenilor mult mai bune. În plus, restricțiile practice și sociale vor împiedica satisfacerea la maximum a tuturor preferințelor simultan, ceea ce înseamnă că roboții trebuie să medieze între preferințele conflictuale – lucru cu care filosofii și cercetătorii sociali se luptă de milenii întregi. Și ce ar trebui să învețe roboții de la oamenii care se bucură de suferința altora? Probabil este mai bine să eliminăm astfel de preferințe din calculele roboților.

Găsirea unei soluții la problema controlării inteligenței artificiale este o sarcină importantă; poate fi, în cuvintele lui Bostrom, „sarcina esențială a vremurilor noastre". Până acum cercetarea inteligenței artificiale s-a concentrat pe sisteme care sunt mai bune la luarea deciziilor, dar acest lucru este diferit de a lua decizii mai bune. Indiferent cât de bună este maximizarea unui algoritm și indiferent cât de precis este modelul său de lume, deciziile unei mașini pot fi inefabil de stupide în ochii unui om obișnuit dacă funcția sa utilitară nu este aliniată cu valorile umane.

Această problemă necesită o modificare a definiției inteligenței artificiale înseși – de la un domeniu care se preocupă de inteligența pură, independentă de obiectiv, la un câmp care se

preocupă cu sisteme care sunt vădit benefice pentru oameni. Dacă luăm în serios această problemă, probabil vom găsi modalități noi de a ne gândi la inteligența artificială, la scopul ei și la relația noastră cu ea.

Capitolul 4
A treia lege

George Dyson

George Dyson este un istoric al științei și tehnologiei și autorul cărților *Baidarka: The Kayak*, *Darwin Among the Machines*, *Project Orion* și *Turing's Cathedral*.

În 2005, George Dyson, istoric al științei și tehnologiei, a vizitat Google la invitația unor ingineri de-acolo. Ocazia a fost reprezentată de cea de-a șaizecea aniversare a propunerii lui John von Neumann de computer digital. În urma acestei vizite, George a scris un eseu, „Catedrala lui Turing", care avertiza publicul pentru prima dată despre ce-și propun să facă fondatorii Google. „Nu scanăm toate aceste cărți pentru a putea fi citite de oameni", îi explica una dintre gazde după prelegere. „Le scanăm pentru a fi citite de o inteligență artificială."

George oferă o poveste alternativă pentru era digitală. Interesele sale au inclus dezvoltarea lui caiacului aleutin, evoluția tehnicii de calcul digitale și a telecomunicațiilor, originile universului digital și o cale nestrăbătută către spațiu. Cariera sa (nu a terminat niciodată liceul, însă a primit un titlu onorific de doctor de la Universitatea

Victoria) s-a dovedit imposibil de clasificat și cărțile sale la fel.

Îi place să evidențieze că tehnica de calcul analogică, despre care se credea cândva că este depășită, la fel ca analizorul diferențial, a revenit. Argumentează că, deși folosim componente digitale, la un anumit moment, tehnica de calcul analogică efectuată de sistem depășește cu mult complexitatea codului digital cu care este construit. Consideră că inteligența artificială adevărată – cu sistemele analogice de control care iau naștere dintr-un substrat digital, la fel cum computerele digitale au luat naștere din componentele analogice după cel de-al Doilea Război Mondial – nu este atât de îndepărtată pe cât credem.

În acest eseu, George contemplă diferența dintre calculele analogice și cele digitale și constată că analogicul este bine și sănătos în lumea contemporană. Răspunsul naturii la încercarea de a programa mașinile să controleze totul poate fi dat sub forma unor mașini fără programare asupra cărora nimeni nu are control.

Istoria tehnicii de calcul poate fi împărțită într-un Vechi Testament și un Nou Testament: înainte și după ce computerele digitale electronice și codurile cărora le-au dat naștere au proliferat pe Pământ. Profeții Vechiului Testament, care ne-au pus la dispoziție logica elementară, i-au inclus pe Thomas Hobbes și Gottfried Wilhelm Leibniz. Profeții Noului Testament i-au inclus pe Alan Turing, John von Neumann, Claude Shannon și Norbert Wiener. Aceștia ne-au pus la dispoziție mașinile.

Alan Turing se întreba ce ar face o mașină inteligentă. John von Neumann se întreba ce ar putea face ca mașinile să se autoreproducă. Claude Shannon se întreba ce ar putea face ca mașinile să comunice fiabil, indiferent de zgomotul care ar interveni. Norbert Wiener se întreba cât timp mai avem până când mașinile vor prelua controlul.

Avertizările lui Wiener cu privire la sistemele de control care se prelungesc dincolo de controlul oamenilor au apărut în 1949, când a fost introdusă prima generație de computere digitale electronice cu programe stocate. Sistemele necesitau o supraveghere directă din partea programatorilor umani, subminând motivele sale de îngrijorare. Care este problema atât timp cât programatorii controlează mașinile? De atunci dezbaterea cu privire la riscurile controlului autonom a rămas asociată cu dezbaterea cu privire la puterile și limitările mașinilor codate digital. În ciuda puterilor lor uluitoare, a fost observată foarte puțină

autonomie reală. Aceasta este o presupunere periculoasă. Dacă puterea tehnicii de calcul digitale este înlocuită de altceva?

Electronica a trecut prin două tranziții fundamentale în ultima sută de ani: de la analogic la digital și de la tuburi vidate la starea solidă. Faptul că aceste tranziții au apărut împreună nu înseamnă că sunt legate inextricabil. La fel cum calculul digital a fost implementat cu ajutorul componentelor cu tuburi vidate, calculul analogic poate fi implementat în starea solidă. Calculul analogic este bine sănătos, chiar dacă tuburile vidate au dispărut din comerț.

Nu există nicio distincție precisă între calculul analogic și cel digital. În general, calculul digital lucrează cu numere întregi, secvențe binare, logică deterministă și timp idealizat în incremente discrete, în timp ce calculul analogic lucrează cu numere reale, logică nedeterministă și funcții continue, inclusiv timpul așa cum există el ca un continuum în lumea reală.

Imaginează-ți că trebuie să găsești mijlocul unui drum. Poți să-i măsori lățimea cu orice increment disponibil și apoi poți calcula digital mijlocul celui mai apropiat increment. Sau poți folosi o sfoară drept computer analogic, mapând lățimea drumului cu lungimea sforii și găsind mijlocul acesteia, fără a fi limitat la incremente, unind capetele sforii și apoi întinzând-o.

Multe sisteme funcționează atât în regim analogic, cât și în regim digital. Un copac integrează o gamă variată de intrări ca funcții continue, dar, dacă tai acel copac, observi că a numărat digital anii tot timpul.

În calculul analogic, complexitatea rezidă în topologia rețelei, și nu în cod. Informația este prelucrată ca funcții continue de valori, precum tensiunea și frecvența relativă a pulsului, și nu prin operații logice pe șiruri discrete de biți. Calculul digital, intolerant la erori sau ambiguități, depinde de corectarea erorilor în orice pas pe care-l face. Calculul analogic tolerează erori, permițându-ți să trăiești cu ele.

Natura folosește codarea digitală pentru depozitarea, replicarea și recombinarea secvențelor de nucleotide, dar se bazează pe

calcul analogic, funcționând pe sisteme nervoase, pentru inteligență și control. Sistemul genetic din toate celulele vii este un computer cu program stocat. Creierele nu sunt.

Computerele digitale execută transformări între două specii de biți: biți ce reprezintă diferențele în spațiu și biți ce reprezintă diferențele în timp. Transformările dintre aceste două forme de informație, secvență și structură, sunt guvernate de programarea computerului și, atât timp cât computerele au nevoie de programatori umani, noi păstrăm controlul.

Computerele analogice mediază și transformările dintre două forme de informații: structura în spațiu și comportamentul în timp. Nu există nici cod, nici programare. Cumva – și nu înțelegem prea bine cum – natura a dezvoltat computere analogice cunoscute sub numele de sisteme nervoase, care înglobează informațiile absorbite din lume. Învață. Unul dintre lucrurile pe care le învață este controlul. Învață să-și controleze propriul comportament și învață să-și controleze mediul în măsura în care pot.

Informatica are o istorie îndelungată – care se întinde chiar înainte de *existența* informaticii – de implementare a rețelelor neuronale, dar majoritatea au fost simulări de rețele neuronale făcute de computere digitale, și nu rețele neuronale așa cum le-a creat natura. Acest lucru începe să se schimbe, de jos în sus, deoarece factorul întreit al războiului cu drone, vehiculelor autonome și telefoanelor celulare împinge dezvoltarea microprocesoarelor neuromorfe care implementează rețele neuronale efective, și nu simulări ale rețelelor neuronale direct în siliciu (și alte substraturi potențiale); și de sus în jos, deoarece cele mai mari și mai de succes întreprinderi ale noastre se întorc din ce în ce mai mult la calculele analogice, încercând să se infiltreze și să controleze lumea.

În timp ce dezbatem inteligența computerelor digitale, calculul digital se furișează tăcut peste cel digital, în același fel în care componentele analogice precum tuburile vidate au fost restructurate pentru a construi computerele digitale în urma celui de-al

Doilea Război Mondial. Procesoarele deterministe individuale în stare finită, ce rulează coduri finite, formează organisme metazoare în stare nonfinită, nedeterministe și la scară largă care zburdă libere în lumea reală. Sistemele hibride analogice și digitale rezultate tratează colectiv fluxurile de biți, la fel cum este fluxul de electroni într-un tub vidat, nu individual, cum sunt tratați biții de dispozitivele în stare discretă care generează fluxul. Biții sunt noii electroni. Analogicul a revenit și natura sa este să preia controlul.

Guvernând totul de la fluxul bunurilor până la fluxul traficului și al ideilor, aceste sisteme funcționează statistic, deoarece informațiile codate cu frecvență de puls sunt prelucrate într-un neuron sau într-un creier. Emergența inteligenței atrage atenția lui *Homo sapiens*, dar ar trebui să ne îngrijorăm mai mult de emergența controlului.

Imaginează-ți că ești în anul 1958 și încerci să aperi partea continentală a Statelor Unite ale Americii de atacurile aeriene. Pentru a distinge aeronavele ostile, unul dintre lucrurile de care ai nevoie, în afară de o rețea de computere și radare cu avertizare în timp util, este o hartă cu traficul aerian comercial, actualizată în timp real. Statele Unite au construit o astfel de hartă și au numit-o SAGE (Semi-Automatic Ground Environment). SAGE a dat la rândul său naștere proiectului Sabre, primul sistem integrat de rezervare a zborurilor în timp real. Sabre și urmașii săi au devenit în curând nu doar o hartă cu locurile disponibile, ci și un sistem care a început să controleze, cu o inteligență descentralizată, unde vor zbura companiile aeriene și când.

Dar nu este oare pe undeva o cameră de control, cu cineva la comenzi? Poate că nu. Să zicem, de exemplu, că vom construi un sistem care să cartografieze traficul de pe autostrăzi în timp real, pur și simplu dând mașinilor acces la acea hartă în schimbul informațiilor despre viteza și localizarea lor la momentul respectiv. Rezultatul este un sistem de control complet descentralizat. Nicăieri nu se află un model de control al sistemului în afara sistemului însuși.

Imaginează-ți că este primul deceniu din secolul al XXI-lea și vrei să urmărești complexitatea relațiilor umane în timp real. Pentru viața socială din cadrul unei facultăți mici, ai putea construi o bază de date centrală, încercând să o ții actualizată, dar întreținerea ei ar deveni copleșitoare dacă sistemul ar fi implementat la o scară mai largă. Mai bine să distribui copii gratuite ale unui cod semiautonom simplu, găzduit local, și să lași rețeaua socială să se actualizeze singură. Codul este executat de computere digitale, dar calculele analogice efectuate de sistem ca întreg depășesc cu mult complexitatea codului pe care îl au la bază. Modelul rezultat, codat cu frecvența pulsului, al graficului social *devine* graficul social. Se întinde prin campus și apoi în lume.

Dar dacă ai vrea să construiești o mașină care să surprindă ce *înseamnă* tot ce e cunoscut de specia umană? Cu Legea lui Moore în spate, nu durează prea mult să digitalizezi toate informațiile din lume. Scanezi toate cărțile tipărite vreodată, colectezi toate e-mailurile scrise vreodată și strângi patruzeci și nouă de ani de imagini video la fiecare douăzeci și patru de ore, în timp ce urmărești unde se află oamenii și ce fac, în timp real. Dar cum surprinzi *sensul*?

Chiar și în epoca tuturor lucrurilor digitale, acesta nu poate fi definit cu o strictețe logică, deoarece sensul, la oameni, nu este fundamental logic. După ce ai colectat toate răspunsurile posibile, nu poți să faci decât să inviți întrebări bine definite și să compilezi o hartă ponderată cu frecvențe de puls pentru felul în care se conectează toate lucrurile între ele. Foarte repede, sistemul tău nu va observa și mapa doar sensul lucrurilor, va începe și să *construiască* sens. În timp, va *controla sensul,* în același fel în care hărțile traficului încep să controleze fluxul traficului chiar dacă nimeni nu pare să fie la control.

Există trei legi ale inteligenței artificiale. Prima, cunoscută sub numele de Legea lui Ashby, după cibernetician W. Ross Ashby, autorul cărții *Design for a Brain*, spune că sistemul de control efectiv trebuie să fie la fel de complex ca sistemul pe care-l controlează.

A doua lege, articulată de John von Neumann, spune că trăsătura definitorie a unui sistem complex este că își constituie singur cea mai simplă descriere comportamentală. Cel mai simplu model complet al unui organism este organismul însuși. Dacă încerci să reduci comportamentul sistemului la o descriere formală, îngreunezi lucrurile, nu le ușurezi.

A treia lege spune că orice sistem suficient de simplu să poată fi înțeles nu va fi suficient de complicat pentru a se comporta inteligent și orice sistem suficient de complicat pentru a se comporta inteligent va fi prea complicat de înțeles.

A treia lege îi liniștește pe cei care cred că, până când vom înțelege inteligența, nu trebuie să ne facem probleme legate de nașterea inteligenței superomenești în mașini. Dar există o portiță în cea de-a treia lege. Este complet posibil să construiești ceva fără să-l înțelegi. Nu trebuie să înțelegi complet cum funcționează un creier pentru a construi unul care funcționează. Aceasta este o portiță pe care nicio supraveghere a algoritmilor de către programatori și consultanții lor etici, oricât ar fi ea de strictă, nu o poate închide. Inteligența artificială ostentativ „bună" este un mit. Relația noastră cu inteligența artificială adevărată va ține întotdeauna de încredere, nu de dovadă.

Ne facem prea multe probleme legate de inteligența mașinilor și nu ne îngrijorăm suficient de autoreproducere, comunicare și control. Următoarea revoluție în puterea de calcul va fi semnalizată de ascensiunea sistemelor analogice, asupra cărora programarea digitală nu mai are niciun control. Răspunsul naturii pentru cei care consideră că pot construi mașini care să controleze totul va fi să le permită să construiască o mașină care-i va controla pe ei în schimb.

Capitolul 5
Ce putem face?

Daniel C. Dennett

Daniel C. Dennett este profesor universitar și profesor Austin B. Fletcher de filosofie și coadministrator al Centrului pentru Studii Cognitive de la Universitatea Tuft. Este autorul a douăsprezece cărți, inclusiv *Consciousness Explained* și, mai recent, *From Bacteria to Bach and Back: the Evolution of Minds.*

Dan Dennett este filosoful preferat al comunității inteligenței artificiale. Este cunoscut în știința cognitivă probabil pentru conceptul său de sisteme intenționale și modelul său pentru conștiința umană, care schițează o arhitectură computațională pentru realizarea fluxului conștiinței în cortexul cerebral puternic paralelizat. Acest computaționalism fără compromisuri a primit critici puternice din partea unor filosofi precum John Searle, David Chalmers și răposatul Jerry Fodor, care au spus că cele mai importante aspecte ale conștiinței – intenționalitatea și *qualia* subiectivă – nu pot fi calculate.

Cu douăzeci și cinci de ani în urmă, îl vizitam pe Marvin Minsky, unul dintre pionierii inteligenței artificiale,

și l-am întrebat de Dan. „Este cel mai bun filosof al nostru din prezent – următorul Bertrand Russell", a spus Marvin, adăugând că, spre deosebire de filosofii tradiționali, Dan studia neuroștiința, lingvistica, inteligența artificială, informatica și psihologia. „Redefinește și reformează rolul filosofului. Bineînțeles, Dan nu înțelege teoria mea despre societatea minții, dar nimeni nu este perfect."

Perspectiva lui Dan asupra eforturilor depuse de cercetătorii din domeniul inteligenței artificiale pentru a crea inteligențe artificiale superinteligente arată întotdeauna o judecată sănătoasă. Ce, să-mi fac probleme eu? În acest eseu, ne reamintește că inteligențele artificiale, mai presus de toate, trebuie privite – și tratate – ca niște instrumente, și nu ca pe niște colegi umani.

A fost interesat de teoria informației încă din zilele sale de studenție de la Oxford. De fapt, mi-a spus că, încă de la începutul carierei sale, era foarte interesat să scrie o carte despre ideile cibernetice ale lui Wiener. În calitate de gânditor care îmbrățișează metoda științifică, din charisma sa face parte și disponibilitatea de a greși. Despre un articol recent intitulat „Ce este informația?", a anunțat: „îl susțin, dar este în revizuire. Deja am trecut dincolo de el și mi-am dat seama că există o cale mai bună de a aborda câteva dintre aceste probleme". Probabil va rămâne calm și rațional cu privire la subiectul cercetării inteligenței artificiale, deși a recunoscut, adesea, că propriile sale idei evoluează – la fel cum ar trebui să facă ideile tuturor.

Mulți au reflectat asupra ironiei de a citi o carte mare când ești prea tânăr să o apreciezi. Să pui un clasic în stiva de cărți *deja citite*, izolându-te astfel de alte influențe din partea lui, spicuind doar câteva idei prost înțelese din el, este o rețetă pentru neglijență care este rareori benignă. Această idee m-a lovit puternic când am citit *Folosirea omului de către om* la peste șaizeci de ani după întâlnirea mea juvenilă cu această carte. Ar trebui să ne facem cu toții un obicei regulat să recitim cărți din tinerețe, unde putem descoperi previzualizări clare ale câtorva dintre „descoperirile" și „invențiile" noastre ulterioare, alături de o bogăție de observații la care eram destinați să fim insensibili până când mințile noastre nu se vor coace, antrena și spori prin confruntarea cu problemele vieții.

Scriind într-o vreme în care tuburile vidate încă erau elementele de fundație ale electronicii și existau doar câteva computere, Norbert Wiener și-a imaginat viitorul pe care-l trăim acum în detalii impresionante și cu destul de puține greșeli evidente. Faimosul articol din 1950 al lui Alan Turing „Mașinile de calcul și inteligența", apărut în revista de filosofie *Mind*, a prevăzut dezvoltarea inteligenței artificiale, la fel a făcut și Wiener, însă Wiener a văzut mai departe și mai adânc, recunoscând că inteligența artificială nu va imita – și înlocui – pur și simplu ființele umane în multe activități inteligente, ci va și schimba oamenii pe parcurs:

„Nu suntem decât vârtejuri într-un râu permanent. Nu suntem ceva care să continue, ci tipare care se perpetuează"*

Când au fost scrise aceste cuvinte, puteau fi respinse cu ușurință ca altă exagerare heracliteană. Da, da, nu poți să pășești în același râu de două ori. Dar conține germenii revoluției în perspectivă. Astăzi știm cum să ne raportăm la sisteme adaptive complexe, atractori bizari, minți extinse și homeostază, o schimbare de perspectivă care promite să șteargă cu buretele „discontinuitatea explicativă"** dintre minte și mecanism, spirit și materie, discontinuitate care este încă apărată cu ferocitate de cartezienii moderni care nu pot suporta ideea că noi – noi *înșine* – suntem tipare de materie purtătoare de informație care se autoperpetuează, și nu „ceva care să continue". Acele tipare sunt remarcabil de rezistente și autorestauratorii, dar în același timp exploatatori versatili, oportuniști și egoiști ai oricărei ocazii noi și disponibile pentru valorificare în goana lor după perpetuare. Și aici devine riscant, spunea Wiener. Când oportunitățile atrăgătoare abundă, suntem deschiși să plătim puțin și să acceptăm costuri de afaceri mici, chiar triviale, pentru a accesa puteri noi. Și devenim rapid atât de dependenți de instrumentele noastre, încât pierdem capacitatea de a prospera fără ele. Opțiunile devin obligatorii.

Este o poveste foarte veche, cu multe capitole bine-cunoscute în istoria evoluționistă. Majoritatea mamiferelor își pot sintetiza propria vitamina C, dar primatele, alegând o dietă ce conține în mare parte fructe, au pierdut această abilitate înnăscută. Suntem acum obligați să ingurgităm vitamina C, dar nu suntem obligați să fim frugivori ca verii noștri, primatele, pentru că am optat pentru o tehnologie care ne permite să facem și să luăm vitaminele atunci când avem nevoie. Tiparele autoperpetuante pe care

---

* *The Human Use of Human Beings* (Boston: Houghton Mifflin, 1954), p. 96 (n.a.).

** Joseph Levine, „Materialism and Qualia: The Explanatory Gap", Pacific Philosophical Quarterly 64 (1983): 354–61 (n.a.).

le numim oameni sunt dependente astăzi de haine, alimente gătite, vitamine, vaccinuri, carduri de credit, telefoane inteligente și internet. Și – mâine, dacă nu deja de astăzi – de inteligența artificială.

Wiener a prevăzut problemele pe care Turing și ceilalți optimiști le-au trecut în mare cu vederea. Pericolul real, a spus el, este

> „ca aceste mașini, deși neajutorate pe cont propriu, să poată fi utilizate de om sau o organizație de oameni pentru a-și spori controlul asupra restului rasei sau că liderii politici ar putea încerca să controleze populațiile nu neapărat prin intermediul mașinilor, ci prin tehnici politice înguste și indiferente la posibilitățile umane, ca și cum ar fi fost, de fapt, concepuți mecanic".*

Puterea, recunoștea el, rezidă în primul rând în algoritmi, nu în componentele hardware pe care rulează, deși hardware-ul de astăzi face posibili algoritmi care ar fi părut ridicol de greoi pe vremea lui Wiener. Ce putem spune despre aceste „tehnici" care sunt „înguste și indiferente la posibilitățile omenești"? Au fost prezentate de multe ori, unele evident benefice, altele evident periculoase, și multe în omniprezentul teren conflictual al controversei.

Gândiți-vă la câteva dintre dispute. Prietenului meu decedat Joe Weizenbaum, succesorul lui Wiener în calitatea de profet Ieremia de la MIT în domeniul hi-tech, îi plăcea să observe despre cardurile de credit că, oricare ar fi punctele lor forte, ofereau de asemenea și o modalitate ieftină și sigură pentru guvern și corporații să urmărească deplasările, obiceiurile și dorințele oamenilor. Anonimitatea numerarului a fost mult subapreciată, în afară de traficanții de droguri și alți infractori, și acum este pe cale de dispariție. Acest lucru poate îngreuna spălarea de bani pe viitor,

* The Human Use of Human Beings (Boston: Houghton Mifflin, 1954), p. 181 (n.a.).

însă programele de căutare ale tiparelor folosite de inteligența artificială contra acestei probleme au efectul secundar de a ne face pe toți mai transparenți la un „bloc de ființe umane" care ar putea „încerca să ne controleze".

În ceea ce privește artele, inovația înregistrărilor digitale audio și video ne permite să plătim un preț mic (în ochii tuturor, dar cel mai arzător, în ochii audiofililor și cinefililor) când abandonăm formatele analogice și, în schimb, oferă o reproducere ușoară – mult prea ușoară? – a operelor de artă cu o fidelitate aproape perfectă. Dar costul ascuns este uriaș. Ministerul Adevărului imaginat de Orwell este acum o posibilitate practică. Tehnici ale inteligenței artificiale de creare a falsurilor aproape nedetectabile pentru „înregistrările" de întâlniri devin acum disponibile, lucru care va face instrumentele de investigare pe care am ajuns să le luăm de-a gata în ultimii 150 de ani să devină depășite. Vom abandona scurta epocă a dovezilor fotografice și ne vom întoarce la lumea de dinainte, a memoriei umane și a încrederii date de standardul aurului, sau vom dezvolta tehnici noi de apărare și atac în cursa pentru înarmare a adevărului? (Ne putem imagina o întoarcere la filmul *analog* expus la lumină, păstrat în sisteme „cu protecție la acces" până când vor fi arătate juriilor etc., dar cât de mult va dura până când cineva găsește o soluție să infecteze aceste sisteme cu îndoială? Una dintre lecțiile grele ale experiențelor recente este că sarcina distrugerii unei reputații pentru credibilitate este mult mai ieftină decât sarcina protejării aceleiași reputații.) Wiener a văzut acest fenomen în cel mai general aspect al său: „pe termen lung, nu mai există nicio diferență între a ne înarma noi înșine sau a ne înarma dușmanii". Era informației este în același timp și era dezinformării.

Ce putem face? Trebuie să ne regândim prioritățile cu ajutorul analizelor eronate, dar pline de pasiune ale lui Wiener, Weizenbaum și ceilalți critici serioși ai tehnofiliei noastre. O frază-cheie mi se pare observația aproape improvizată de mai sus că „aceste mașini sunt aproape neajutorate pe cont propriu". După cum am argumentat recent, construim instrumente, nu colegi, și marele

pericol este să nu apreciem diferența, pe care ar trebui să încercăm să o evidențiem, marcând-o și apărând-o cu inovații politice și legale.

Poate că cea mai bună cale de a vedea ce ratăm este să ne aducem aminte că Alan Turing însuși a suferit de o lipsă de imaginație complet de înțeles când a formulat faimosul său test Turing. După cum știe toată lumea, este o adaptare a „jocului său de imitație" în care un om, care nu se vede, comunică verbal cu un judecător, încercă să-l convingă că este, de fapt, o femeie, în timp ce o femeie, de asemenea ascunsă și care comunică cu judecătorul, încearcă să-l convingă pe judecător că ea este femeia. Turing credea că aceasta va fi o provocare grea pentru un bărbat (sau pentru o femeie care se preface că este bărbat), exploatând o bogăție de cunoștințe despre cum gândește și reacționează celălalt sex, despre ce are tendința să favorizeze și să ignore. Cu siguranță (*hopa!*) orice bărbat care poate învinge o femeie la a fi perceput ca fiind femeie ar fi un agent inteligent. Ce nu a prevăzut Turing este puterea inteligenței artificiale cu învățare profundă de a obține această bogăție de informații într-o formă exploatabilă *fără să fie nevoie să o înțeleagă*. Turing s-a gândit la un agent perspicace și plin de imaginație (așadar, conștient) care-și gândea isteț răspunsurile în baza „teoriei" detaliate despre ce este probabil să facă și să spună o femeie. Design inteligent descendent, pe scurt. Cu siguranță nu s-a gândit că un bărbat care câștigă jocul imitației va *deveni* cumva femeie; își imagina că va exista tot timpul conștiința bărbatului la baza acestui spectacol. Premisa ascunsă din argumentul lui Turing era: doar un *agent* conștient inteligent poate dezvolta și controla o strategie victorioasă în jocul imitației. Așadar, era convingător pentru Turing (și pentru alții, printre care mă număr și eu, un dârz apărător al testului lui Turing) să argumenteze că „o mașină de calcul" care ar putea trece drept om într-o întrecere cu un om nu este neapărat conștientă așa cum este omul, dar trebuie să fie un soi de *agent* conștient. Cred că este o poziție care se poate apăra – singura poziție care se poate apăra –, dar trebuie să înțelegi cât de perspicace și de ingenios ar

trebui să fie un judecător pentru a expune superficialitatea fațadei pe care ar putea-o prezenta o inteligență artificială cu învățare profundă (un instrument, nu un coleg).

Ce nu a prevăzut Turing este capacitatea nemaivăzută a computerelor foarte rapide de a cerne fără discernământ prin seturile mari de date, pe care internetul le pune la dispoziție în cantități nesfârșite, găsind tipare probabilistice în activitățile oamenilor care pot fi utilizate pentru a improviza răspunsuri care să pară „autentice" în dispozitivul de ieșire pentru aproape orice probă pe care un judecător s-ar gândi s-o administreze. Și Wiener a subestimat această posibilitate, ținând cont de faptul că slăbiciunea care trădează mașina este să nu fie capabilă să

> „ia în considerare vastul interval de probabilitate ce caracterizează situația umană"*

Dar a lua în considerare acest vast interval de probabilitate este exact punctul forte al inteligenței artificiale. Singura slăbiciune din armura inteligenței artificiale este acest cuvânt „vast"; posibilitățile omenești, grație limbajului și culturii pe care le produce, sunt cu adevărat vaste**. Indiferent de cât de multe tipare am putea găsi cu inteligența artificială în fluviul de date care se găsesc pe internet, există mult mai multe posibilități decât sunt înregistrate aici. Doar o fracție (dar nu o fracție nesemnificativă) din înțelepciunea, proiectele, repertoriile și prostioarele acumulate ale lumii au ajuns pe internet, dar, probabil, o tactică mai bună pentru judecător când confruntă un candidat pentru testul Turing este nu să *caute* aceste elemente, ci să le *creeze* de la zero.

---

* The Human Use of Human Beings (Boston: Houghton Mifflin, 1954), p. 181 (n.a.).

** În cartea *Darwin's Dangerous Idea*, 1995, p. 109, am definit versiunea capitalizată, „vast" însemnând *mult mai mult decât astronomic*, și noțiunea complementară „efemer" pentru a înlocui obișnuitele exagerări *infinit* și *infinitezimal* pentru discutarea acelor posibilități care nu sunt oficial infinite, însă infinite în toate scopurile practice (n.a.).

Inteligența artificială în manifestările sale curente este parazitară pe inteligența umană. Se îndoapă nediferențiat cu tot ce a fost produs de creatorii umani și extrage tiparele care pot fi extrase de acolo – inclusiv obiceiurile noastre dăunătoare*. Aceste mașini nu au (încă) obiectivele, strategiile sau capacitățile de autocritică și inovație care să le permită să transceandă bazele lor de date gândindu-se reflexiv la propria lor gândire și la propriile lor obiective. Sunt, după cum spune Wiener, neajutorate, nu în sensul de agenți *încătușați* sau agenți *cu capacități reduse*, ci în sensul că nu sunt agenți deloc – nu au capacitatea de a fi „puse în mișcare de motivele" (cum spunea Kant) care li se prezintă. Este important să păstrăm lucrurile astfel, ceea ce va necesita muncă.

Una dintre erorile din cartea lui Weizenbaum, *Computer Power and Human Reason*, de care am încercat să-l conving în multe ore de discuții, era că nu se putea hotărî niciodată ce să apere: *inteligența artificială este imposibilă!* Sau *inteligența artificială este posibilă, dar malefică!* Voia să argumenteze, împreună cu John Searle și Roger Penrose, că „inteligența artificială puternică" este imposibilă, dar nu există argumente solide care să justifice această concluzie. La urma urmei, tot ce știm acum sugerează că, după cum am mai spus, suntem roboți făcuți din roboți făcuți din roboți... până la nivelul proteinelor motorii și altor elemente asemănătoare, fără ingrediente magice aruncate în amestec pe parcurs. Mesajul mai important și mai defensibil al lui Weizenbaum era că nu ar trebui să ne străduim să creăm inteligență artificială puternică și ar trebui să fim foarte atenți în ceea ce privește sistemele de inteligență artificială pe care le putem crea și pe care le-am creat deja. După cum te-ai putea aștepta, teza defensibilă este un hibrid: *inteligența artificială* (inteligența artificială puternică) *este posibilă în principiu, dar nu este de dorit. Inteligența artificială care este posibilă din punct*

* Aylin Caliskan-Islam, Joanna J. Bryson și Arvind Narayanan, „Semantics Derived Automatically from Language Corpora Contain Human- Like Biases", Science 356, nr. 6334 (14 aprilie, 2017): 183– 86, DOI:10.1126/science.aal4230 (n.a.).

de vedere practic nu este neapărat rea decât dacă se confundă cu inteligența artificială puternică!

Intervalul dintre sistemele de astăzi și sistemele science-fiction ce domină imaginația populară este încă uriaș, deși mulți oameni – atât nespecialiști, cât și specialiști – reușesc să-l subestimeze. Să ne gândim la Watson, de la IBM, care poate fi luat drept un reper de încredere pentru imaginațiile noastre deocamdată. Este rezultatul unui proiect de cercetare și dezvoltare la scară foarte largă, întinzându-se peste multe secole pe persoană, după cum notează George Church în aceste pagini, folosește de mii de ori mai multă energie decât un creier omenesc (o limitare tehnologică ce, după cum observă tot el, poate fi temporară). Victoria sa în *Jeopardy!* a fost un triumf veritabil, făcut posibil de restricțiile formale ale regulilor emisiunii, dar, pentru a putea concura, chiar și aceste reguli au trebuit să fie revizuite (un compromis: renunți la puțină versatilitate, la puțină umanitate și obții un spectacol care mulțumește mulțimea). Watson nu este o companie plăcută, în ciuda reclamelor amăgitoare de la IBM care sugerează o capacitate conversațională generală, iar transformarea lui Watson într-un *agent* multidimensional plauzibil ar fi echivalentul transformării unui calculator de buzunar în Watson. Watson ar putea fi o facultate de bază utilă pentru un astfel de agent, dar mai asemănător unui cerebel sau unei amigdale decât unei minți – în cel mai bun caz, un subsistem specializat care poate juca un rol de suport mare, dar nici măcar aproape să poată formula scopuri, planuri și structuri cu perspicacitate în baza experiențelor sale conversaționale.

De ce am vrea să-l transformăm pe Watson într-un agent gânditor și creativ? Poate că ideea genială a lui Turing de test operativ ne-a dus într-o capcană: sarcina de a crea cel puțin iluzia unei persoane reale în spatele ecranului, care să ofere o punte peste „valea misterioasă". Pericolul aici este că, de când a lansat Turing această provocare – care era, la urma urmei, o provocare menită să-i *păcălească pe* judecători –, creatorii inteligenței artificiale au încercat să peticească valea cu artificii omenești drăgălașe, efecte de Disneyficare care-i vor încânta și dezarma pe neinițiați.

ELIZA lui Weizenbaum a fost exemplul principal pentru o astfel de iluzionare superficială și disperarea sa la ușurința cu care programul său rizibil de simplu și superficial îi putea convinge pe oameni că au o conversație candidă este cea care l-a făcut să-și asume această misiune.

Și motivele sale de îngrijorare erau întemeiate. Dacă am învățat ceva din competițiile restricționate ale testului Turing pentru Premiul Loebner este că până și oamenii inteligenți care nu sunt familiari cu posibilitățile și scurtăturile programării pe calculator sunt păcăliți cu ușurință de trucuri simple. Atitudinile oamenilor din domeniul inteligenței artificiale față de aceste metode de a disimula în „interfața de utilizator" au mers de la dispreț la celebrare, cu o apreciere generală că aceste trucuri nu sunt profunde, dar pot fi puternice. O schimbare de atitudine foarte binevenită este recunoașterea candidă că ornamentele omenești sunt *reclamă falsă* – ceva de condamnat, nu de aplaudat.

Cum am putea realiza acest lucru? Când vom recunoaște că oamenii încep să ia decizii de viață și de moarte în mare parte la „sfatul" sistemelor de inteligență artificială ale căror operații interne sunt de nepătruns în practică, putem vedea un motiv suficient pentru cei care-i încurajează pe oameni în orice fel să acorde mai multă încredere acestor sisteme decât ar trebui să fie trași la răspundere atât moral, cât și legal. Sistemele de inteligență artificială sunt instrumente foarte puternice – atât de puternice încât până și experții au motive întemeiate să nu aibă încredere în propria judecată privind „judecățile" propriilor lor instrumente. Însă, dacă acești utilizatori de instrumente urmează să beneficieze, financiar sau altfel, de folosirea lor în *terra incognita*, trebuie să se asigure că o pot face responsabil, cu un control și o justificare maxime. Licențierea și alocarea operatorilor, la fel cum îi licențiem pe farmaciști (și pe operatorii de macarale!) și alți specialiști ale căror greșeli și erori de judecată pot avea consecințe grave, îi pot, din cauza presiunii companiilor de asigurări și a altor asigurători, obliga pe creatorii sistemelor de inteligență artificială să facă eforturi

extraordinare să caute și să arate slăbiciuni și lipsuri în produsele lor și să-i instruiască pe cei care au dreptul să le folosească.

Ne putem imagina un fel de test Turing invers, în care judecătorul este supus procesului; până când acesta poate detecta slăbiciunea, limitele depășite, lipsurile dintr-un sistem, nicio licență de operare nu va fi emisă. Instruirea mentală necesară pentru a obține certificarea de judecător va fi dificilă. Îndemnul de a adopta o poziție intențională, tactica noastră normală ori de câte ori întâlnim ceea ce pare a fi un agent inteligent, este aproape copleșitor de puternic. Într-adevăr, capacitatea de a rezista tentației de a trata ceea ce pare a fi o persoană ca pe o persoană este un talent urât, care duhnește a rasism sau specism. Mulți oameni ar considera cultivarea unei abordări atât de sceptice respingătoare din punct de vedere moral și putem anticipa că până și cei mai specializați utilizatori de sistem vor cădea pradă ocazional tentației de a se „împrieteni" cu instrumentele lor, fie și numai pentru a-și alina disconfortul provocat de realizarea îndatoririlor pe care le au. Indiferent cât de lipsiți de scrupule ar fi designerii inteligenței artificiale în deghizarea „umană" a bunurilor produse, ne putem aștepta ca obiceiuri noi de gândire, inițiative și artificii conversaționale, capcane și blufuri să ia naștere în acest context nou al acțiunii umane. Listele hilar de lungi cu efecte secundare cunoscute ale medicamentelor noi care au reclame la televizor vor fi umbrite de revelațiile obligatorii ale întrebărilor care *nu pot* primi răspunsuri responsabile din partea unor anumite sisteme, cu pedepse dure pentru cei care „scapă" defecte în produsele lor. Este recunoscut la scară largă că o parte considerabilă din inechitatea economică în creștere de astăzi se datorează bogăției acumulate de antreprenorii digitali; trebuie să introducem legi care limitează buzunarele lor adânci pentru binele public. Câteva dintre cele mai adânci buzunare sunt în mod voluntar în opoziție cu aceste obligații de a servi societatea în primul rând și de a câștiga bani în al doilea rând, dar nu ar trebui să ne bizuim doar pe buna credință.

Nu avem nevoie de agenți conștienți artificiali. Există o abundență de agenți conștienți naturali, suficienți să administreze orice sarcină rezervată acestor entități speciale și privilegiate. Avem nevoie de instrumente inteligente. Instrumentele nu au drepturi și nu ar trebui să aibă sentimente care pot fi rănite sau să poată să răspundă cu resentimente la „abuzurile" la care sunt supuse de către utilizatorii inepți*. Unul dintre motivele pentru care nu ar trebui să facem agenți conștienți artificiali este că, oricât de autonomi ar deveni (și, în principiu, pot deveni la fel de autonomi, de perfectibili și de creativi ca orice persoană), aceștia nu ar împărtăși cu noi – fără instrucțiuni speciale, care pot fi anulate –, agenții conștienți naturali, vulnerabilitățile și mortalitatea noastră.

I-am provocat la un moment dat pe studenții de la un seminar de la Tufts pe care l-am ținut împreună cu Matthias Scheutz pe tema agenților artificiali și autonomiei: dați-mi specificațiile unui robot care ar putea semna un contract cu putere juridică pentru voi – nu ca surogat pentru un proprietar uman, ci pe cont propriu. Problema nu este să-l faci să înțeleagă clauzele sau să-l faci să miște un stilou pe o bucată de hârtie, ci de a avea și de a *merita* statutul juridic de agent moral responsabil. Copiii nu pot semna astfel de contracte, și nici persoanele cu dizabilități al căror statut juridic le cere să fie sub îngrijirea sau responsabilitatea unui tutore sau a altui fel de responsabil. Problema pentru roboții care ar vrea să obțină un astfel de statut solemn este că, la fel ca Superman, sunt prea invulnerabili pentru a putea face o promisiune credibilă. Dacă nu și-ar respecta obligațiile asumate, ce s-ar întâmpla? Care ar fi pedeapsa pentru încălcarea promisiunilor? Să fie închiși într-o celulă sau, mai probabil, să fie demontați? Să fie închisă nu ar fi o mare problemă pentru o inteligență artificială decât dacă

* Joanna J. Bryson, „Robots Should Be Slaves", în Close Engagement with Artificial Companions, Yorick Wilks, ed. (Amsterdam, The Netherlands: John Benjamins, 2010), 63– 74; http://www.cs.bath.ac.uk/~jjb/ftp/Bryson-Slaves-Joanna J. Bryson, „Patiency Is Not a Virtue: AI and the Design of Ethical Systems", https://www.cs.bath.ac.uk/~jjb/ftp/Bryson-Patiency-AAAISS16.pdf (n.a.).

instalăm în prealabil plăcerea de a călători care nu poate fi ignorată sau dezactivată de inteligența artificială pe cont propriu (și ar fi sistematic foarte dificil să implementăm o astfel de soluție cu succes dată fiind istețimea și cunoașterea de sine presupuse ale inteligenței artificiale); iar demontarea unui sistem de inteligență artificială (fie el robot sau agent imobilizat la pat cum este Watson) nu înseamnă să-l omori dacă informațiile stocate în designul și în softul sale sunt păstrate. Ușurința înregistrărilor și transmiterilor digitale – inovația care le permite softurilor și datelor să fie practic nemuritoare – scoate roboții din lumea vulnerabilului (cel puțin roboții pe care ni-i putem imagina de obicei, cu softuri și memorii digitale). Dacă nu este deja evident, gândește-te cum ar fi afectată moralitatea omenească dacă am putea să facem „cópii de rezervă" pentru oameni în fiecare săptămână, de exemplu. Te arunci sâmbătă cu capul în jos de pe un pod înalt fără coardă și ai un iureș de trăiri pe care nu ți-l vei aduce aminte duminică dimineață, când se activează copia de rezervă făcută vineri seară, dar te poți bucura de clipul video al morții tale.

Așadar, ceea ce creăm nu este sau nu ar trebui să fie agenți conștienți, umanoizi, ci o entitate cu totul nouă, asemenea oracolelor, fără conștiință, fără teamă de moarte, fără iubire și ură care să o distragă, fără personalitate (dar cu tot soiul de metehne și ciudățenii care vor fi identificate fără îndoială drept „personalitatea" sistemului): cutii de adevăruri (dacă suntem norocoși) aproape sigur contaminate cu o difuzie de falsuri. Va fi suficient de greu să învățăm să trăim cu ele, fără să ne mai torturăm cu fantezii despre singularitate în care aceste inteligențe artificiale ne vor subjuga, la propriu. Folosirea omului *de către om* va fi schimbată în curând – încă o dată – pentru totdeauna, dar putem să prindem bine cârma și să navigăm printre câteva dintre pericole dacă ne asumăm responsabilitatea traiectoriei.

Capitolul 6
Debandada inumană în care ne-au băgat mașinile

Rodney Brooks

Rodney Brooks este informatician, profesor Panasonic de robotică, emerit, MIT, fostul administrator al Laboratorului de Inteligență artificială de la MIT și al Laboratorului de Informatică și Inteligență Artificială de la MIT. Este autorul cărții *Flesh and Machines*.

Roboticianul Rodney Brooks, care a apărut în documentul lui Errol Morris din 1997 intitulat *Fast, Cheap and Out of Control* alături de un dresor de lei, un horticultor și un expert în șobolanii-cârtiță golași, a fost descris de un cronicar ca fiind „zâmbitor cu o sclipire de nebunie în ochi". Dar acest lucru se poate spune despre aproape toți vizionarii.

Câțiva ani mai târziu în cariera sa, Brooks, așa cum se cade unuia dintre cei mai mari roboticieni din lume, a sugerat că „supraantropomorfizăm oamenii, care sunt, la urma urmei, niște simple mașini". A continuat prin prezentarea unei viziuni plăcute a unei lumi viitoare a inteligențelor artificiale în care „distincția dintre noi și roboți va dispărea". A recunoscut, de asemenea, și că are o perspectivă împărțită asupra lumii. „Asemenea unui

cercetător religios, păstrez două seturi de credințe incongruente și acționez asupra fiecăruia în diverse circumstanțe", scria el. „Consider că această transcendență între sistemele de credințe va permite omenirii să accepte în cele din urmă roboții ca mașini emoționale și apoi să înceapă să empatizeze cu ei și să le confere liber-arbitru, respect și, în cele din urmă, drepturi."

Asta se întâmpla în 2002. În aceste pagini, abordează o perspectivă mai caustică, fie ea mai restrânsă; este alarmat de măsura în care am ajuns să ne bazăm pe sistemele generalizate, care nu sunt numai exploatatoare, ci și vulnerabile, ca urmare a dezvoltării mult prea rapide a ingineriei software – progres ce pare că a întrecut impunerea siguranțelor eficiente și fiabile.

Matematicienii și oamenii de știință sunt adesea limitați în modul în care percep imaginea de ansamblu, dincolo de domeniul lor de activitate, din cauza instrumentelor și metaforelor pe care le folosesc în munca lor. Norbert Wiener nu este o excepție, și cred că nici eu.

Când a scris *Folosirea omului de către om*, Wiener era la sfârșitul erei înțelegerii mașinilor și animalelor ca simple procese fizice și la începutul erei noastre curente de înțelegere a mașinilor și animalelor ca procese computaționale. Bănuiesc că vor fi ere viitoare ale căror instrumente vor fi la fel de diferite de instrumentele celor două ere la intersecția cărora s-a aflat Wiener pe cât de diferite sunt instrumentele celor două ere între ele.

Wiener a fost un titan al erei anterioare și s-a folosit de instrumentele inventate încă de pe vremea lui Newton și Leibniz pentru a descrie și a analiza procesele continue din lumea fizică. În 1948, a publicat *Cibernetica*, un cuvânt pe care l-a născocit pentru a descrie știința comunicării și controlului atât la mașini, cât și la animale. Astăzi ne referim la ideile din această carte drept teoria controlului, o disciplină indispensabilă pentru proiectarea și analiza mașinilor fizice, neglijând în mare parte spusele lui Wiener despre știința comunicării. Inovațiile lui Wiener au avut la bază în mare parte munca sa din timpul celui de-al Doilea Război Mondial cu mecanisme de țintire și arme de foc antiaeriene. A adus rigoare matematică în proiectarea tehnologiei ale cărei procese de proiectare erau în mare parte euristice: de la

sistemele de alimentare cu apă ale romanilor până la motorul cu aburi al lui Watt și dezvoltarea primelor automobile.

Ne putem imagina o versiune contingentă diferită a istoriei noastre intelectuale și tehnologice dacă nu ar fi apărut în scenă Alan Turing și John von Neumann, ambii aducând contribuții majore fundației proceselor de calcul. Turing a contribuit cu un model de calcul fundamental – cunoscut acum ca mașina Turing – în lucrarea sa *Despre numerele computabile cu aplicare pe problema deciziei**, scrisă și revizuită în 1936 și publicată în 1937. În aceste mașini, o bandă liniară de simboluri dintr-un alfabet finit codează intrarea pentru o problemă de calcul, punând la dispoziție și mediul de lucru pentru calcul. Era necesară o mașină diferită pentru fiecare problemă de calcul; lucrările ulterioare scrise de alți cercetători vor arăta că pe una dintre mașini, cunoscută drept Mașina Turing Universală, un set arbitrar de instrucțiuni de calcul poate fi codat pe aceeași bandă.

În anii 1940, von Neumann a dezvoltat o mașină abstractă cu autoreproducere denumită automatul celular. În acest caz, ocupa un subset finit dintr-o matrice bidimensională infinită de pătrate, fiecare conținând un singur simbol dintr-un alfabet finit cu douăzeci și nouă de simboluri diferite – restul matricei infinite începe goală. Simbolurile individuale din fiecare pătrat se schimbă la unison în baza unei reguli complexe, dar finite, aplicată simbolului curent din pătrat și vecinilor săi imediați. Sub regula complexă pe care a dezvoltat-o von Neumann, majoritatea simbolurilor din majoritatea pătratelor rămân neschimbate și câteva se schimbă în fiecare fază. Prin urmare, dacă cineva privește pătratele care nu sunt goale, pare că este vorba despre o structură constantă în care se desfășoară o mică activitate. Când mașina abstractă a lui von Neumann se reproducea, făcea o copie a sa în altă regiune a planului. În cadrul „mașinii" se afla o linie orizontală de pătrate care acționa ca o bandă liniară finită, utilizând un subset din

* În original, „On Computable Numbers with an Application to the Entscheidungsproblem" (n.t.).

alfabetul finit. Simbolurile din acele pătrate codau mașina din care făceau parte. În timpul reproducerii mașinii, „banda" se putea deplasa la stânga sau la dreapta și era interpretată (transcrisă) ca instrucțiuni (conversie) pentru noua „mașină" care era construită, apoi copiată (replicată) –, iar noua copie era pusă în interiorul noii mașini pentru reproducere ulterioară. Francis Crick și James Watson au arătat mai târziu, în 1953, cum o astfel de bandă poate fi particularizată în biologie de o moleculă lungă de ADN cu alfabetul său finit format din patru baze azotate: guanină, citozină, adenină și timină (G, C, A și T)*. La fel ca în cazul mașinii lui von Neumann, în reproducerea biologică, secvența liniară de simboluri din ADN este interpretată – prin transcriere în molecule ARN, care sunt apoi convertite în proteine, structurile care creează o celulă nouă – și apoi ADN-ul este replicat și încapsulat în noua celulă.

A doua lucrare fundamentală a fost în 1945, într-un raport „versiune preliminară" privind designul unui computer digital, unde Neumann susținea o memorie care putea conține atât instrucțiuni, cât și date**. Acum este cunoscut sub numele de computer cu arhitectură Neumann – diferit de computerele cu arhitectură Harvard în care există două memorii separate, una pentru instrucțiuni și una pentru date. Vasta majoritate de procesoare pentru computere construite în epoca Legii lui Moore sunt bazate pe arhitectura von Neumann, inclusiv cele care stau la baza centrelor noastre de date, laptopurilor și telefoanelor noastre inteligente. Arhitectura de computer digital a lui von Neumann este conceptual aceeași generalizare – de la computerele digitale timpurii construite cu relee electromagnetice la Universitatea Harvard și Bletchley Park – care apare când treci de la mașină

* „A Structure for Deoxyribose Nucleic Acid", Nature 171 (1953): 737–38 (n.a.).

** J. von Neumann, „First Draft of a Report on the EDVAC", IEEE Annals of the History of Computing 15 (1993): 27– 75. Von Neumann este listat ca fiind singurul autor, în timp ce alții au contribuit la conceptele pe care le-a prezentat el; astfel întreaga arhitectură i-a fost atribuită lui (n.a.).

Turing specială la o mașină Turing universală. În plus, automatele sale cu replicare automată împărtășesc o similitudine fundamentală atât cu construcția unei mașini Turing, cât și cu mecanismele celulelor biologice care se reproduc pe bază de ADN. Chiar și în ziua de astăzi mai există dezbateri între intelectuali dacă von Neumann a văzut interconexiunile dintre aceste trei lucrări, a lui Turing și cele două ale sale. Revizuirea făcută de Turing lucrării sale a fost făcută când el și von Neuman erau ambii la Princeton; într-adevăr, după ce și-a obținut doctoratul, Turing aproape că a rămas în calitate de postdoctorand al lui Neumann.

Fără Turing și von Neumann, cibernetica lui Wiener ar fi putut rămâne un mod dominant de gândire și de structurare a tehnologiei mult mai mult timp decât scurta sa perioadă de supremație. În această versiune imaginară de istorie, am fi putut trăi astăzi într-o lume de fapt *steam-punk*, nu doar să observăm particularizările sale fantastice la Maker Faires!

Ce vreau să spun este că Wiener privea lumea – fizică, biologică și (în *sens uman*) sociologică – într-un mod foarte special. A analizat lumea ca un set continuu de variabile, după cum explică în capitolul 1 împreună cu o apreciere a termodinamicii prin lentila statisticilor lui Gibbs. De asemenea, încearcă să introducă forțat un model slab și neconvingător al informației ca un mesaj care se transmite între entități fizice și biologice. Pentru mine și din punctul de vedere al informațiilor de astăzi, după șaptezeci de ani, instrumentele sale par cumplit de inadecvate pentru descrierea mecanismelor care stau la baza sistemelor biologice, iar astfel nu a înțeles cum mecanisme similare ar putea până la urmă să fie încorporate în sisteme de calcul tehnologice – așa cum sunt încorporate acum. Tehnologiile dominante de astăzi au fost dezvoltate în lumea lui Turing și von Neumann, mai degrabă decât în lumea lui Wiener.

În prima revoluție industrială, energia de la un motor cu abur sau de o roată de moară a fost utilizată de muncitorii umani pentru a înlocui propria lor energie. În loc să fie o sursă de energie pentru munca fizică, oamenii au devenit modulatorii a cum va fi

utilizată o sursă puternică de energie. Dar, din cauza faptului că motoarele cu abur și roțile de moară trebuiau să fie mari pentru a fi o utilizare rentabilă a capitalului și pentru că în secolul al XVIII-lea singura tehnologie pentru distribuirea spațială a energiei era mecanică și funcționa doar pe o rază foarte mică, mulți muncitori erau nevoiți să se înghesuie în jurul sursei de energie. Wiener argumentează corect că abilitatea de a transmite energie ca electricitate a determinat o a doua revoluție industrială. Acum sursa de energie poate fi la o distanță mare de locul de utilizare și, de la începutul secolului XX, producția a putut fi mult mai dispersată după construirea rețelelor de distribuție a energiei electrice.

Wiener argumentează mai departe că încă o tehnologie nouă, cea mașinilor de calcul care luau naștere în vremea sa, va da naștere încă unei revoluții. Mașinile despre care vorbește par să fie atât analoge, cât și (probabil) digitale; și evidențiază, în *The Human Use of Human Beings*, faptul că, de vreme ce vor putea lua decizii, atât muncitorii din producție, cât și cei de la birouri ar putea fi reduși la a fi simple rotițe într-o mașinărie mult mai mare. Se temea că oamenii s-ar putea folosi și abuza unii pe alții prin structuri organizaționale pe care această capacitate le va încuraja. Cu siguranță, am trăit acest scenariu în ultimii șaizeci de ani și această perturbare nici măcar nu se apropie de final.

Însă perspectiva sa bazată pe fizică în ceea ce privește calculul îl proteja de realizarea a cât de rău ar putea sta lucrurile cu adevărat. Vedea capacitatea mașinilor de a comunica drept furnizarea unei modalități noi și mai inumane de a exercita comanda și controlul. Nu a întrevăzut faptul că, în câteva decenii, sistemele de calcul vor deveni mai asemănătoare sistemelor biologice și se pare, din descrierile sale din capitolul 10 din propria sa lucrare despre modelarea anumitor aspecte ale biologiei, că a subapreciat grav multele ordine de magnitudine de complexitate ulterioare ale biologiei peste cele ale fizicii. Astăzi suntem într-o situație mult mai complexă decât a prevăzut el și mă tem că este mult mai periculos decât în cele mai sumbre predicții ale sale.

În anii 1960, calculul a început să se bazeze ferm pe fundația stabilită de Turing și de von Neumann și aceasta era reprezentată de calculul digital, bazat pe ideea unor alfabete finite pe care le-au utilizat ambii. O secvență de o lungime arbitrară sau un *string*, formată din caracterele unui alfabet finit poate fi codată ca un număr întreg unic. Așa cum se întâmplă și cu mașinile Turing, formalismul calculului a devenit formalismul calculării unei funcții cu valori întregi pentru o singură intrare evaluată ca număr întreg.

Turing și von Neumann au murit ambii în anii 1950 și, la momentul respectiv, așa vedeau ei calculul. Niciunul dintre ei nu a prevăzut creșterea exponențială a puterii de calcul pe care urma să o aducă Legea lui Moore – și nici cât de universale urmau să devină mașinile de calcul. Nu a prevăzut nici dezvoltarea a două modele pentru sistemele noastre de calcul, fiecare dintre ele prezentând o mare amenințare pentru societatea umană.

Primul își are rădăcina în abstracțiile adoptate. În cei cincizeci de ani de cursă alimentată de Legea lui Moore de a produce software care să exploateze dublarea capacității computerelor la fiecare doi ani, grija și certificarea disciplinelor de inginerie tipice au încetat să mai fie folosite. Dezvoltarea softurilor era rapidă și supusă erorilor. Această dezvoltare rapidă a softurilor fără standarde de corectitudine a deschis multe rute de exploatare a arhitecturii Neumann de stocare a datelor și instrucțiunilor în aceeași memorie. Una dintre cele mai comune rute, cunoscută sub numele de „depășire a memoriei *buffer*“*, implică un număr de intrare (sau un șir lung de caractere) care este mai mare decât a prevăzut programatorul și depășește memoria revărsându-se în zona în care sunt stocate instrucțiunile. Dacă proiectează cu atenție un număr de intrare care este mult prea mare, cineva care folosește un soft îl poate infecta cu instrucțiuni care nu sunt prevăzute de programator, modificând astfel ceea ce face. Acestea sunt bazele creării unui virus de computer – numit astfel datorită similitudinii

* În original, „buffer overrun“. Microsoft Language Portal oferă și varianta de traducere „depășire a tamponului“ (n.t.).

sale cu un virus biologic. Cel din urmă injectează ADN suplimentar într-o celulă, iar mecanismul de transcriere și traducere al celulei respective îl interpretează orbește făcând proteine care pot fi dăunătoare celulei-gazdă. În plus, mecanismul de replicare al celulei are grijă de multiplicarea virusului. Astfel o entitate străină mică poate prelua controlul asupra unei entități mult mai mari și-i poate modifica comportamentul în moduri neprevăzute.

Acestea și alte forme de atacuri digitale ne-au privat de securitatea vieților noastre de zi cu zi. Ne bazăm pe computere pentru aproape orice acum. Ne bazăm pe computere pentru infrastructura noastră de electricitate, gaze, drumuri, mașini, trenuri și avioane; acestea sunt toate vulnerabile. Ne bazăm pe computere pentru *banking*, plata facturilor, conturile noastre de economii, ipotecile, achizițiile de bunuri și servicii – și acestea sunt toate vulnerabile. Ne bazăm pe computere pentru distracție, comunicații personale și profesionale, securitatea noastră fizică acasă, informațiile noastre despre lume și sistemele noastre de vot – toate vulnerabile. Niciuna dintre acestea nu va fi reparată prea curând. Între timp, multe aspecte ale societății noastre sunt deschise unor atacuri violente, fie venite din partea unor infractori liber-profesioniști, fie din partea adversarilor statului.

A doua dezvoltare este că calculul a trecut de simplele funcții de calcul. În schimb, programele rămân online continuu, putând aduna astfel date despre o secvență de interogări. În schema Wiener / Turing / von Neumann, putem privi tiparul comunicării pentru un browser web drept:

Utilizator: dă-mi pagina web A.

Browser: iată pagina web A.

Utilizator: dă-mi pagina web B.

Browser: iată pagina web B.

Acum, în schimb, poate arăta așa:

Utilizator: dă-mi pagina web A.

Browser: iată pagina web A. (Și voi reține în secret că mi-ai cerut pagina web A.)

Utilizator: dă-mi pagina web B.

Browser: iată pagina web B. Văd o corelație între conținutul ei și cel al cererii de mai devreme, pagina web A, așa că-mi voi actualiza modelul pe care îl am pentru tine, utilizatorul, și-l voi transmite companiei care m-a produs.

Când mașina nu mai calculează pur și simplu o funcție, ci întreține în schimb o stare, poate începe să facă inferențe despre om în funcție de secvența de cereri care i-a fost prezentată. Și când diverse programe corelează mai multe fluxuri de cereri – de exemplu, corelează căutările de pagini web cu postările social-media sau plata serviciilor pe altă platformă sau timpul pe care l-ai petrecut vizionând o reclamă sau unde a fost condus utilizatorul cu GPS-ul telefonului inteligent –, sistemele totale ale mai multor programe ce comunică între ele și cu diverse baze de date conduc la o cu totul nouă pierdere a confidențialității. Marele salt exploator făcut de atât de multe companii de pe Coasta de Vest a fost să monetizeze aceste inferențe fără permisiunea conștientă a persoanei care generează interacțiunile cu platformele mașinilor de calcul.

Wiener, Turing și von Neumann nu au putut prevedea complexitatea acestor platforme, în timp ce bălmăjeala în termeni juridici din contractele condițiilor de utilizare pe care oamenii la acceptă voit, fără să aibă cea mai mică idee ce presupun acestea, îi face să renunțe la drepturi la care altfel nici nu s-ar gândi să renunțe într-o interacțiune unu la unu cu o persoană. Platformele de calcul au devenit un scut în spatele căruia anumite companii

se ascund pentru a-i exploata inuman pe alții. În alte țări, guvernele sunt cele care fac aceste manipulări și acolo obiectivul nu este profitul, ci suprimarea disidenței.

Omenirea s-a pus într-o situație dificilă: suntem exploatați de companii care, paradoxal, ne oferă servicii pe care ni le dorim și, în același timp, viețile noastre depind de multe sisteme bazate pe softuri care sunt deschise atacurilor. Să ieșim din această încurcătură este un proiect pe termen lung. Va necesita planificare, legi și, cel mai important, *leadership* moral. *Leadership*-ul moral este prima și cea mai mare provocare.

Capitolul 7
Unitatea inteligenței

Frank Wilczek

Frank Wilczek este profesor Herman Feshbach de fizică la MIT, laureat al Premiului Nobel pentru fizică în 2004 și autorul cărții *A Beautiful Question: Finding Nature's Deep Design.*

L-am cunoscut pe Frank Wilczek în anii 1980, când m-a invitat acasă la el, în Princeton, pentru a discuta despre anioni. „Adresa este strada Mercer, nr. 112", mi-a scris el. „Caută o casă fără străduță de parcare." Și iată-mă după câteva ore în vechea cameră de zi a lui Einstein, discutând cu un viitor laureat al Premiului Nobel în fizică. Dacă Frank era la fel de impresionat ca mine de împrejurimi, nu o să ghiciți niciodată. Singurul său comentariu era despre dificultatea de a găsi un loc de parcare în fața unei „case fără străduță de parcare".

Spre deosebire de majoritatea fizicienilor teoreticieni, Frank era interesat de mult timp de inteligența artificială, mărturie stând aceste trei „observații":

1. „Francis Crick i-a spus „ipoteza uimitoare": și anume, conștiința, cunoscută și ca minte, este o proprietate emergentă a materiei", afirmație care, dacă este adevărată, arată că „toată inteligența este inteligență automată. Ce diferențiază inteligența naturală de cea artificială nu este natura ei, ci modul în care este creată".
2. „Inteligența artificială nu este produsul unei invazii a extratereștrilor. Este un artefact al unei culturi omenești speciale și reflectă valorile acelei culturi."
3. „Afirmația șocantă a lui David Hume «rațiunea este și trebuie să fie sclava pasiunilor» a fost scrisă în 1783 și trebuia să se aplice, bineînțeles, numai rațiunii și pasiunilor omenești... Dar ideea logică/filosofică a lui Hume rămâne valabilă și pentru inteligența artificială. Mai simplu: stimulentele, și nu logica abstractă, stau la baza comportamentului."

Observă că „marea poveste a secolelor 20 și 21 este că, odată cu dezvoltarea computerelor, învățăm să calculăm consecințele legilor fundamentale din ce în ce mai bine. Există și un ciclu de *feedback*: când înțelegi materia mai bine, poți proiecta computere mai bune, care-ți vor permite să calculezi mai bine. Este un fel de spirală ascendentă"

Aici, argumentează că inteligența umană deține avantajul, deocamdată – însă viitorul nostru, nelegat de sistemul nostru solar și fără îndoială nici de galaxia noastră, nu se va înfăptui niciodată fără ajutorul inteligențelor noastre artificiale.

I. Un răspuns simplu la întrebări contencioase

- Poate o inteligență artificială să fie conștientă?
- Poate o inteligență artificială să fie creativă?
- Poate o inteligență artificială să fie malefică?

Aceste întrebări sunt puse adesea astăzi, atât în mass-media populară, cât și în dezbaterile științifice informate. Dar discuțiile nu par să fie niciodată convergente. Aici voi începe prin a oferi răspunsuri după cum urmează:

În baza psihologiei fiziologice, neurobiologiei și fizicii, ar fi surprinzător dacă răspunsurile nu ar fi da, da și da. Motivul este simplu și totuși profund: dovezile din domeniile respective fac foarte posibil faptul că nu există nicio împărțire clară între inteligența naturală și cea artificială.

În cartea sa din 1994 pe această temă, renumitul biolog Francis Crick a propus o „ipoteză uluitoare": că mintea ia naștere din materie. A spus că mintea, în toate aspectele sale, „nu este nimic mai mult decât un ansamblu de celule nervoase și moleculele lor asociate".

„Ipoteza uluitoare" este de fapt fundamentarea neuroștiinței moderne. Oamenii încearcă să înțeleagă cum funcționează mințile înțelegând cum funcționează creierul și încearcă să înțeleagă cum funcționează creierul studiind cum este codată informația în semnale electrice și chimice, transformate de procese fizice și cum

este utilizată pentru a controla comportamentul. În această încercare științifică, nu permit comportamente extrafizice. Până acum, în mii de experimente minunate, această strategie nu a dat greș niciodată. Nu s-a dovedit niciodată necesar să permitem influenței conștiinței sau creativității eliberate de activitatea creierului să explice un eveniment psihofizic sau neurobiologic observat. Nimeni nu a întâlnit vreodată o putere a minții care să fie separată de evenimente fizice convenționale din organismele biologice. Deși sunt multe lucruri pe care nu le înțelegem cu privire la creier și minte, „ipoteza uluitoare" a rămas intactă.

Dacă ne lărgim perspectiva dincolo de neurobiologie și luăm în considerare întregul spectru de experimente științifice, cazul devine și mai fascinant. În fizica modernă, centrele de interes sunt adesea fenomene foarte delicate. Pentru a le investiga, experimentatorii trebuie să ia multe măsuri de precauție împotriva contaminării cu „zgomot". Adesea li se pare necesar să construiască protecție suplimentară contra electricității statice și câmpurilor magnetice; să compenseze pentru vibrații mărunte datorate microcutremurelor sau mașinilor în trecere; să funcționeze la temperaturi extrem de mici și în vid cu presiune; și așa mai departe. Dar există o excepție notabilă: nu li s-a părut niciodată necesar să acorde permisiuni pentru ce gândesc oamenii din apropiere (sau din depărtare). Nu par să existe „unde mentale", diferite de procesele fizice cunoscute, capabile să influențeze evenimentele fizice.

Acea concluzie, luată la valoare nominală, șterge distincția dintre inteligența naturală și artificială. Implică faptul că, dacă ar fi să duplicăm sau să simulăm cu precizie procesele fizice care apar în creier – așa cum, în principiu, putem – și să conectăm intrările și ieșirile la organe de simț și la mușchi, atunci am reproduce, într-un artefact fizic, manifestările observate ale inteligenței naturale. Nu ar lipsi nimic observabil. Ca observator, nu aș avea motive în plus sau în minus să presupun conștiință, creativitate sau răutate acelui artefact pentru care nu atribui acele proprietăți omologilor săi naturali, cum ar fi alte ființe umane.

Astfel, combinând „ipoteza uluitoare" a lui Crick din neurobiologie cu dovezi puternice din fizică, deducem că inteligența naturală este un caz special de inteligență artificială. Această concluzie merită un nume și o voi numi „corolarul uluitor".

Cu ea avem răspunsul la cele trei întrebări ale noastre. De vreme ce conștiința, creativitatea și răutatea sunt trăsături evidente ale inteligenței naturale omenești, sunt și trăsături posibile ale inteligenței artificiale.

Cu o sută de ani în urmă sau chiar cincizeci, să crezi că mintea ia naștere din materie și să inferezi corolarul nostru că inteligența naturală este un caz special de inteligență artificială, ar fi presupus un salt de credință. În lumina multelor lipsuri conexe – abisuri, de fapt – din înțelegerea contemporană a biologiei și fizicii, acestea erau propuneri cu adevărat îndoielnice. Dar dezvoltările epocale din acele domenii au schimbat imaginea:

În biologie: cu un secol în urmă, nu numai gândurile, ci și metabolismul, ereditatea și percepția erau aspecte profund misterioase ale vieții care sfidau explicațiile fizice. Astăzi, bineînțeles, avem relatări extrem de bogate și detaliate despre metabolism, ereditate și multe aspecte ale percepției, de jos în sus, începând de la nivelul molecular.

În fizică: după un secol de fizică cuantică și aplicarea ei pe materiale, fizicienii au descoperit, în repetate rânduri, cât de bogat și ciudat poate fi comportamentul materiei. Superconductorii, laserele și multe alte minuni demonstrează că ansamblele mari de unități moleculare, fiecare simplu în sine, pot prezenta comportamente „emergente", calitativ noi, rămânând complet obediente legilor fizicii. Chimia, inclusiv biochimia, este cornul abundenței în ceea ce privește fenomenele emergente, toate fiind acum înrădăcinate puternic în fizică. Fizicianul deschizător de drumuri Philip Anderson, într-un eseu intitulat „Mai mult este diferit", prezintă o discuție clasică despre emergență. Începe prin a recunoaște că „ipoteza reducționistă [adică, integralitatea explicațiilor fizice bazată pe interacțiunile cunoscute ale părților simple] poate fi o simplă controversă în rândurile filosofilor, dar consider că este

acceptată fără echivoc de vasta majoritate a oamenilor de știință activi". Însă continuă prin a evidenția că „comportamentul agregatelor mari și complexe de particule elementare, se pare, nu trebuie înțeles în termenii unei simple extrapolări a proprietăților câtorva particule". Fiecare nivel nou de mărime și complexitate suportă forme noi de organizare, ale căror tipare codează informații în moduri noi și al căror comportament este cel mai bine descris în concepte noi.

Computerele electronice sunt exemple magnifice de emergență. Aici toate cărțile sunt pe față. Inginerii proiectează cu regularitate, de jos în sus, în baza unor principii fizice cunoscute (și destul de sofisticate), mașini care prelucrează informațiile în moduri extrem de impresionante. iPhone-ul tău te bate la șah, adună și îți pune la dispoziție rapid informații despre orice și mai face și niște poze minunate. Deoarece procesul prin care computerele, telefoanele inteligente și alte obiecte inteligente sunt produse este complet transparent, nu încape îndoială despre capacitățile lor minunate că iau naștere din procese fizice obișnuite, pe care le putem urmări până la nivelul electronilor, fotonilor, quarcilor și gluonilor. Evident, materia brută poate deveni destul de inteligentă.

Dați-mi voie să rezum argumentul. Din două ipoteze cu sprijin puternic, am tras o concluzie simplă:

- Mintea omenească ia naștere din materie.
- Materia este definită de fizică.
- *Așadar*, mintea umană ia naștere din procese fizice pe care le înțelegem și pe care le putem reproduce artificial.
- *Așadar*, inteligența naturală este un caz special de inteligență artificială.

Bineînțeles, „corolarul nostru uluitor" poate da greș; primele două linii ale acestui argument sunt ipoteze. Dar eroarea lor ar

* *Science* 177, nr. 4047 (4 august, 1972): 393–96 (n.a.).

conduce la o descoperire fundamentală – un fenomen semnificativ nou, cu consecințe fizice la scară largă, care are loc în circumstanțe fizice neremarcabile și bine studiate (de exemplu, materialele, temperaturile și presiunile din creierele omenești), însă care a reușit cumva să se sustragă multe decenii arsenalului sofisticat de instrumente al investigatorilor. O astfel de descoperire ar fi... uluitoare.

II. Viitorul inteligenței

Face parte din natura umană să-și îmbunătățească mintea și corpul. Istoric, vestimentația, ochelarii și ceasurile sunt exemple de augmentări din ce în ce mai sofisticate care ne îmbunătățesc rezistența, percepția și conștiința. Sunt îmbunătățiri majore ale zestrei umane naturale a căror familiaritate nu trebuie să ne orbească în ce privește profunzimea lor. Astăzi telefoanele inteligente și internetul aduc dorința umană de augmentare în domenii mai centrale identității noastre de ființe inteligente. Ne conferă, în consecință, acces rapid la o conștiință colectivă vastă și la memorie colectivă vastă.

În același timp, inteligențe artificiale autonome au devenit campioni mondiali într-o gamă variată de jocuri „cerebrale", cum ar fi șahul și Go, și au preluat controlul asupra sarcinilor sofisticate de recunoaștere a tiparelor, cum ar fi reconstituirea întâmplărilor din timpul reacțiilor complexe ce au loc în Large Hadron Collider, de la furtuna de particule emergente până la găsirea unor particule noi; sau colectarea indiciilor dintr-o radiografie neclară, un RMN funcțional și alte tipuri de imagini pentru diagnosticarea problemelor medicale.

Unde ne duce această dorință de autoîmbunătățire și inovație? Deși secvența exactă de evenimente și scara de timp pe care se desfășoară sunt imposibil de preconizat (sau, cel puțin, pe mine mă depășesc), câteva idei de bază sugerează că, în cele din urmă, cele mai puternice întrupări ale minții vor fi foarte diferite de creierele umane așa cum le cunoaștem noi astăzi.

Să ne gândim la șase factori prin care tehnologiile de procesare a informației depășesc capacitățile umane – vast, calitativ, sau ambele:

- *Viteză:* mișcarea orchestrată a electronilor, care este inima procesării moderne a informației artificiale, poate fi mult mai rapidă decât procesele de difuzie și schimbare chimică prin care funcționează creierele. Frecvențele procesoarelor moderne se apropie de 10 GHz, ceea ce înseamnă 10 miliarde de operații pe secundă. Nicio măsură a vitezei nu se aplică varietății năucitoare a proceselor creierului, dar o limitare fundamentală este latența potențialelor de acțiune, care le limitează spațierea la câteva zeci pe secundă. Probabil nu este niciun accident că „frecvența cadrelor" la care putem distinge că filmele sunt de fapt secvențe de imagini este de aproximativ 40 pe secundă. Prin urmare, procesarea electronică este de aproape un miliard de ori mai rapidă.
- *Dimensiune:* dimensiunea liniară a unui neuron tipic este de aproximativ 10 microni. Dimensiunile moleculare, care stabilesc o limită practică, sunt de aproximativ 10 000 de ori mai mici și unitățile de procesare artificiale se apropie de această scară. Dimensiunile reduse fac comunicarea mai eficientă.
- *Stabilitate:* în timp ce memoria umană este esențial continuă (analogă), memoria artificială poate încorpora caracteristici (digitale) discrete. În timp ce cantitățile analoge se pot eroda, cantitățile digitale pot fi stocate, împrospătate și întreținute cu o acuratețe completă.
- *Ciclul de lucru:* creierele omenești obosesc de la efort. Au nevoie de pauze pentru a se hrăni și pentru a dormi. Poartă povara îmbătrânirii. Și, cel mai important: mor.
- *Modularitate (arhitectură deschisă):* deoarece procesoarele de inteligență artificială pot accepta interfețe digitale precis definite, pot asimila cu ușurință modele noi. Astfel, dacă

vrem ca un computer să „vadă" ultraviolete sau infraroșii sau să „audă" ultrasunete, putem alimenta ieșire direct din senzorul corespunzător în „sistemul său nervos". Arhitectura creierelor este mult mai închisă și opacă, iar sistemul imunitar uman rezistă activ implanturilor.

- *Promptitudine cuantică:* un caz de modularitate merită o mențiune specială datorită potențialului său pe termen lung. Recent, fizicienii și specialiștii în informații au ajuns să aprecieze faptul că principiile mecanicii cuantice suportă noi principii de calcul, care pot crea forme calitative noi de prelucrare a informației și (plauzibil) niveluri noi de inteligență. Dar aceste posibilități se bazează pe aspectele comportamentului cuantic care sunt foarte delicate și par să fie necorespunzătoare pentru o interfață cu mediul cald, umed și haotic al creierelor umane.

Evident, ca platforme pentru inteligență, creierele umane sunt departe de a fi optime. Totuși, deși roboții de curățenie versatili sau soldații mecanici își găsesc piețe rentabile, în prezent nu există nicio mașină care să se apropie de tipul de inteligență umană cu scopuri generale de care ar avea nevoie acele aplicații. În ciuda slăbiciunii lor relative pe multe fronturi, creierele omenești au avantaje uriașe față de competiția lor artificială. Dați-mi voie să menționez cinci:

- *Tridimensionalitatea:* deși, așa cum am observat deja, dimensiunile liniare ale unităților de procesare artificiale existente sunt mult mai mici decât cele ale creierelor, procedura prin care sunt create – centrată pe litografie (în principiu, schițare) – este esențialmente bidimensională. Acest lucru se vede clar în geometria plăcilor și cipurilor de computer. Bineînțeles, se pot conecta mai multe plăci, dar spațierea dintre straturi este mult mai mare și comunicarea mult mai puțin eficientă decât în cadrul aceluiași strat. Creierele folosesc mai bine cele trei dimensiuni.

- *Autoreparația:* creierele omenești își pot reveni sau pot ocoli multe tipuri de vătămări sau erori. Computerele trebuie adesea să fie reparate sau repornite extern.
- *Conectivitatea:* neuronii umani suportă câteva sute de conexiuni (sinapse). În plus, tiparul complex al acestor conexiuni este plin de sens. (Vezi următoarea mea idee.) Unitățile computerelor fac doar câteva conexiuni, în tipare regulate și fixe.
- *Dezvoltarea* (asamblare automată cu sculptură interactivă): creierul uman își crește unitățile prin diviziune celulară și le orchestrează în structuri coerente prin mișcare și înfășurare. De asemenea proliferează o abundență de conexiuni între celule. O parte importantă a sculptării sale apare prin procese active în timpul copilăriei sale, când individul interacționează cu mediul său. În acest proces, multe conexiuni sunt eliminate, în timp ce altele sunt fortificate, în funcție de eficacitatea utilizării. Astfel structura fină a creierului este reglată prin interacțiunea cu lumea externă – o sursă bogată de informații și *feedback!*
- *Integrarea* (senzori și activatori)*:* creierul omenesc vine echipat cu o varietate de organe senzoriale, printre care notabili sunt ochii concrescuți și cu mecanisme versatile de control, inclusiv mâini care construiesc, picioare care merg și guri care vorbesc. Acești senzori și mecanisme de control sunt integrate perfect în centrele de procesare a informației din creier, fiind perfecționate de-a lungul a milioane de ani de selecție naturală. Interpretăm semnalele lor brute și controlăm acțiunile lor la scară mare cu o atenție conștientă minimă. Reversul medaliei este că nu știm cum facem asta, și implementarea este opacă. Se dovedește a fi surprinzător de dificil să ajungem la standarde umane în ceea ce privește aceste funcții de intrare-ieșire „de rutină".

Aceste avantaje ale creierelor omenești asupra artefactelor proiectate în prezent sunt profunde. Creierele omenești oferă o dovadă de existență înălțătoare, arătându-ne câteva metode de a

ieși și mai mult din materie. Când va ajunge ingineria la acest nivel, dacă va ajunge vreodată?

Nu știu sigur, dar dați-mi voie să vă spun câteva păreri informate. Provocările tridimensionalității și, într-o mai mică măsură, ale autoreparației nu par covârșitoare. Pun niște probleme inginerești dificile, dar multe îmbunătățiri incrementale sunt ușor de imaginat și se întrevăd căi clare de înaintare. Și, deși puterile ochilor omenești, ale mâinilor și ale altor organe senzoriale și mecanisme de acționare sunt incredibil de eficiente, capacitățile lor sunt departe de a depăși orice limite fizice. Sistemele optice pot face poze cu o rezoluție mai mare în spațiu, timp și culoare și în mai multe regiuni ale spectrului electromagnetic; roboții se pot mișca mai repede și pot fi mai puternici; și așa mai departe. În aceste domenii, componentele necesare pentru performanțe supraomenești, pe mai multe axe, sunt deja disponibile. Dificultatea rezidă în a introduce și a scoate rapid informația din ele, într-o limbă a unităților de prelucrare a informației.

Și astfel ajungem la restul avantajelor – și, cred eu, cele mai profunde – creierelor asupra dispozitivelor artificiale, care iau naștere din conectivitatea lor și din dezvoltarea interactivă. Aceste două avantaje sunt sinergetice, deoarece dezvoltarea interactivă sculptează structura puternic conectată și vastă a creierului de copil, activată de creșteri exponențiale de neuroni și sinapse, pentru a se modela în instrumentul extraordinar care urmează să devină. Informaticienii încep să descopere puterea arhitecturii creierului: rețelele neuronale, al căror design fundamental, așa cum sugerează și numele, a fost inspirat direct de designul creierului, au avut succese spectaculoase în domeniul jocurilor și în recunoașterea tiparelor, așa cum s-a observat. Dar ingineria de astăzi nu are nimic comparabil – în domeniul (în prezent) ezoteric al mașinilor cu reproducere automată – cu puterea și versatilitatea neuronilor și sinapselor lor. Aceasta poate deveni o frontieră nouă și minunată a cercetării. Și aici biologia ne poate deschide drumul, pe măsură ce înțelegem dezvoltarea biologică suficient de bine pentru a-i imita esența.

În ansamblu, avantajele inteligenței artificiale față de inteligența naturală par să fie permanente, iar avantajele inteligenței naturale față de inteligența artificială, deși sunt substanțiale în prezent, par trecătoare. Cred că vor mai trece multe decenii până când ingineria le va prinde din urmă, dar – dacă nu vor apărea războaie catastrofale, schimbări climatice bruște sau molime, astfel încât progresul tehnologic să poată rămâne viguros – puține secole.

Dacă estimarea este corectă, putem anticipa câteva generații în care oamenii, împuterniciți și augmentați de dispozitive inteligente, vor coexista cu inteligențele artificiale autonome din ce în ce mai capabile, fiind astfel supuse unei evoluții rapide. Date fiind avantajele inerente pe care le vor oferi dispozitivele proiectate în cele din urmă, avangarda acestei evoluții va fi reprezentată de cyborgi și superminți, și nu de niște *homo sapiens* ornați.

Un alt elan important va veni din explorarea mediilor ostile, atât de pe Pământ (de exemplu, din adâncul oceanelor), cât și, în special, din spațiu. Corpul omenesc nu este adaptat la condițiile care se află în afara unui interval restrâns de temperaturi, presiuni și compoziții atmosferice. Are nevoie de o varietate mare de nutrienți specifici și complecși, precum și de multă apă. De asemenea, nu este obișnuit cu radiațiile. După cum a demonstrat clar programul spațial, este dificil și scump să ții oamenii în afara zonei lor de confort terestru. Cyborgii sau inteligențele artificiale autonome se pot dovedi mult mai eficienți în aceste explorări. Inteligențele artificiale cuantice, cu sensibilitatea lor la zgomot, ar putea fi chiar mai fericite în răceala și întunericul spațiului interstelar.

Într-un pasaj impresionant din romanul *Odd John* (1935), geniul singular al genului science fiction, Olaf Stapledon are un erou, o inteligență (mutantă) supraomenească, ce descrie specia *homo sapiens* ca „Arheopterixul spiritului". Îi spune acest lucru cu căldură prietenului și biografului său, care este un om obișnuit. Arheopterix a fost o creatură nobilă și o punte către creaturi mai măreț.

Capitolul 8
Să aspirăm la mai mult decât doar să nu devenim depășiți

Max Tegmark

Max Tegmarck este fizician MIT și cercetător în domeniul inteligenței artificiale, președintele Institutului Future for Life, director științific al Institutului Foundational Questions și autorul cărților *Our Mathematical Universe* și *Life 3.0: Being Human in the Age of Artificial Intelligence.*

L-am cunoscut pe Max Tegmark cu câțiva ani în urmă, când mi l-a prezentat colegul său de la MIT, Alan Guth, părintele teoriei inflației. Fizician teoretician și cosmolog distins el însuși, grija principală a lui Max în prezent este riscul existențial care stă la pândă implicat de crearea unei inteligențe artificiale generale (adică, una care rivalizează inteligenței umane). Cu patru ani în urmă, Max a înființat, împreună cu Jaan Tallinn și alții, Institutul Future of Life, care se prezintă ca „o organizație deschisă ce lucrează pentru a se asigura că cele mai puternice tehnologii ale viitorului sunt benefice umanității". În timpul turneului de prezentare a cărții sale în Londra, era în focurile planificării pentru Institutul Future of Life și recunoaște că a fost

mișcat până la lacrimi într-o stație de metrou după o vizită la Muzeul de Științe din Londra, cu expoziții pe tema realizărilor tehnologice ale umanității. Tot acest progres impresionant este în zadar? Consiliul științific al Institutului Future of Life îi include pe Elon Musk, Frank Wilczek, George Church, Stuart Russell și pe filosoful de la Oxford Nick Bostrom, cel care a conceput un *Gedankenexperiment** citat adesea, ce are ca urmare o lume plină de agrafe de hârtie și nimic altceva, produsă de o inteligență artificială generală (aparent) binevoitoare, care nu făcea decât să respecte comenzile primite. Institutul sponsorizează conferințe (Porto Rico 2015, Asilomar 2017) pe tema problemelor de siguranță create de inteligența artificială și, în 2018, a instituit o competiție pentru burse ce se concentrează pe cercetarea maximizării beneficiilor sociale ale inteligenței artificiale generale.

Deși Max este uneori listat – de necunoscători – în tabăra panicarzilor, el crede, la fel ca și Frank Wilczek, într-un viitor care va beneficia enorm de inteligența artificială generală dacă, în încercarea de a o crea, putem să menținem pentru aceasta specia umană ca prioritate.

* „Experiment mental" în limba germană (n.t.).

Deși există o mare controversă despre momentul în care inteligența artificială va avea un impact semnificativ asupra umanității, situația este mai clară din perspectivă cosmică: viața de pe Pământ care dezvoltă tehnologie se grăbește să devină depășită fără să se gândească în mod serios la consecințe. Acest lucru mi se pare jenant, dat fiind faptul că putem crea oportunități uimitoare astfel încât umanitatea să prospere mai mult decât oricând, dacă îndrăznim să-i dăm un curs mai ambițios.

Universul nostru a devenit conștient de sine după 13,8 miliarde de ani de la nașterea sa. Pe o planetă mică și albastră, părți minuscule de Univers au descoperit că ceea ce au crezut ei cândva că este întreaga existență este o parte infimă din ceva mult mai mare: un sistem solar într-o galaxie într-un univers cu peste 100 de miliarde de alte galaxii, aranjate într-un tipar elaborat de grupuri, clustere și superclustere.

Conștiința este trezirea cosmică; a transformat Universul nostru dintr-un zombi fără minte și fără conștiință de sine într-un ecosistem viu ce adăpostește reflecție de sine, frumusețe, speranță, sens și scop. Dacă această trezire nu ar fi avut loc niciodată, Universul nostru nu ar fi avut niciun sens – un spațiu uriaș irosit. Dacă Universul nostru se va stinge permanent din cauza unei calamități cosmice sau a unei nenorociri create de noi înșine, va deveni din nou fără sens.

Pe de altă parte, lucrurile s-ar putea chiar îmbunătăți. Nu știm încă dacă noi, oamenii, suntem singurii privitori de stele din

cosmos, nici măcar dacă suntem primii, dar am aflat deja suficiente lucruri despre Univers pentru a ști că are potențialul de a se trezi mult mai deplin decât s-a trezit până acum. Pionieri din domeniul inteligenței artificiale precum Norbert Wiener ne-au învățat că o trezire suplimentară a capacității Universului nostru de a procesa și experimenta informații nu necesită neapărat eoni de evoluție suplimentară, ci poate că doar câteva decenii de ingeniozitate științifică omenească.

Am putea fi ca acea primă sclipire de conștiință de sine pe care ai trăit-o când te-ai trezit din somn azi de dimineață, o premoniție a conștiinței mult superioare ca va urma când deschizi ochii și te trezești de-a binelea. Poate că superinteligența artificială va permite vieții să se infiltreze în tot cosmosul și să prospere miliarde sau trilioane de ani, și poate că acest lucru se va întâmpla datorită unor decizii pe care le vom lua aici, pe planeta noastră, în decursul vieții noastre.

Sau poate că umanitatea va pieri printr-o calamitate determinată de ea însăși, cauzată de puterea tehnologiei care crește mai repede decât înțelepciunea cu care o administrăm.

Dezbaterea în desfășurare cu privire la impactul inteligenței artificiale asupra societății

Mulți gânditori resping ideea de superinteligență ca fiind science-fiction, deoarece văd inteligența ca pe ceva misterios ce poate exista doar în organisme biologice – în special în oameni – și ca fiind fundamental limitată la ceea ce pot face oamenii de astăzi. Dar, din perspectiva mea de fizician, inteligența nu este decât un tip de prelucrare a informației realizat de particule elementare care se mișcă, și nu există nicio lege a fizicii care să spună că cineva nu poate construi mașini mai inteligente decât noi din orice punct de vedere, putând astfel să însămânțeze viața cosmică. Acest lucru sugerează că am văzut doar vârful aisbergului inteligenței; există un potențial uimitor pentru deblocarea întregii

inteligențe latente din natură și de a o utiliza pentru a ajuta umanitatea să înflorească sau să se ofilească.

Alții, printre care se numără și câțiva dintre autorii acestui volum, nu cred în construirea unei inteligențe artificiale generale (entitate ce poate efectua orice sarcină cognitivă cel puțin la fel de bine ca oamenii), nu pentru că o consideră fizic imposibilă, ci pentru că o consideră prea dificil de obținut pentru oameni în mai puțin de un secol. Printre cercetătorii profesioniști din domeniul inteligenței artificiale, ambele tipuri de respingere au devenit perspective minoritare datorită descoperirilor recente. Mulți oameni se așteaptă ca inteligența artificială generală să fie obținută într-un secol, iar prognoza medie este în doar câteva decenii. Un studiu recent al cercetătorilor din domeniul inteligenței artificiale realizat de Vincent Müller și Nick Bostrom conchide:

> Rezultatele arată părerea experților conform căreia sistemele de inteligență artificială probabil (peste 50%) vor egala capacitățile generale ale omului până în anii 2040–2050 și foarte probabil (cu o probabilitate de 90%) până în anul 2075. De la atingerea capacităților omenești, se va ajunge la superinteligență în 2 ani (10%) până la 30 de ani (75%) ulterior*.

În perspectiva cosmică de giga-ani, nu contează prea mult dacă inteligența artificială generală sosește în 30 sau în 300 de ani, așa că haideți să ne concentrăm pe consecințe, și nu pe timp.

În primul rând, noi, oamenii, am descoperit cum să replicăm anumite procese naturale cu mașinile, făcându-ne propria căldură, lumină și putere mecanică. Gradual, am observat că și corpurile noastre sunt mașini și descoperirea celulelor nervoase a începeșat

---

* Vincent C. Müller and Nick Bostrom, „Future Progress in Artificial Intelligence: A Survey of Expert Opinion", în Fundamental Issues of Artificial Intelligence, ed. Vincent C. Müller (Elveția: Springer International Publishing, 2016), 555– 72, https://nickbostrom.com/papers/survey.pdf (n.a.).

granițele dintre corp și minte. În cele din urmă, am început să construim mașini care să ne întrească atât mușchii, cât și mințile. În prezent, suntem eclipsați de mașini în realizarea multor sarcini cognitive restrânse, de la memorare și aritmetică până la jocuri, și suntem pe cale să fim întrecuți în multe alte privințe, de la condus până la investiții și diagnosticare medicală. În cazul în care comunitatea inteligenței artificiale reușește să-și atingă obiectivul său inițial de a construi o inteligență artificială generală, atunci vom fi, prin definiție, eclipsați în ceea ce privește toate sarcinile noastre cognitive.

Acest lucru naște multe întrebări evidente. De exemplu, cine va controla inteligența artificială generală va controla și Pământul? Ar trebui să încercăm să controlăm mașinile superinteligente? Dacă nu, ne putem asigura că înțeleg, adoptă și păstrează valorile umane? Așa cum spunea Norbert Wiener în *The Human Use of Human Beings*:

> „Vai de noi dacă lăsăm [mașina] să ne decidă comportamentul înainte de a examina legile acțiunilor sale și de a ști bine dacă acest comportament al mașinii va fi ghidat de principii acceptabile pentru noi! Pe de altă parte, mașina [...] care poate învăța și lua decizii în baza a ceea ce a învățat, nu va fi obligată în niciun fel să ia decizii pe care noi ar fi trebuit să le luăm sau care să fie acceptabile pentru noi".

Și cine suntem „noi" în această ecuație? Cine stabilește dacă „aceste decizii sunt acceptabile"? Chiar dacă puterile viitoare decid să-i ajute pe oameni să supraviețuiască și să prospere, cum vom găsi sens și scop în viețile noastre dacă nu mai este nevoie de noi pentru nimic?

Dezbaterea cu privire la impactul inteligenței artificiale asupra societății s-a schimbat semnificativ în ultimii câțiva ani. În 2014, puținele discuții despre riscurile inteligenței artificiale aveau tendința să fie catalogate drept încercări luddite de creare a fricii, pentru unul dintre următoarele două motive incompatibile:

1. Inteligența artificială generală era prezentată exagerat și nu urma să apară timp de cel puțin încă un secol.
2. Inteligența artificială generală putea să apară mai devreme, dar aproape garantat va fi benefică.

Astăzi discuțiile despre impactul inteligenței artificiale asupra societății sunt peste tot și activitățile ce vizează siguranța și etica inteligenței artificiale au fost mutate în companii, universități și conferințe academice. Poziția controversată a cercetării siguranței inteligenței artificiale nu mai înseamnă susținerea, ci negarea ei. Întrucât scrisoarea deschisă care a reieșit dintr-o conferință despre inteligența artificială ce a avut loc în Porto Rico în anul 2015 (și a ajutat la sporirea siguranței inteligenței artificiale) vorbea doar în termeni vagi despre importanța de a păstra benefică inteligența artificială, principiile Asilomar privind inteligența artificială (vezi mai jos) aveau dinți ascuțiți: menționează explicit autoîmbunătățirea, superinteligența și riscul existențial, fiind semnate de lideri din domeniul inteligenței artificiale și de peste o mie de cercetători din domeniu din întreaga lume.

Cu toate acestea, cele mai multe discuții sunt limitate la impactul pe termen scurt al inteligenței artificiale concentrate, iar comunitatea mai largă acordă doar o atenție limitată transformărilor dramatice pe care inteligența artificială generală le poate aduce în curând vieții de pe Pământ. De ce?

De ce ne grăbim să devenim depășiți și de ce evităm să vorbim despre asta

În primul rând, din motive simple de economie. De câte ori găsim o cale nouă de a face munca omenească depășită, construind mașini care o fac mai bine și mai ieftin, cea mai mare parte a societății câștigă: cei care construiesc și utilizează mașinile fac profit, iar consumatorii obțin produse mai ieftine. Acest lucru va fi la fel de adevărat și pentru inteligențele artificiale generale cu rol de investitor sau inteligențele artificiale generale cu rol de

cercetător, cum a fost și pentru mașinile de țesut, excavatoare și roboții industriali. În trecut, angajații scoși de pe piața muncii și-au găsit alte meserii, dar acest stimulent economic elementar va rămâne chiar dacă nu mai este cazul. Existența inteligențelor artificiale generale ieftine înseamnă, prin definiție, că *toate* meseriile vor putea fi făcute mai ieftin de mașini, deci oricine pretinde că „oamenii vor găsi întotdeauna meserii noi și bine plătite" pretinde, de fapt, că cercetătorii din domeniul inteligenței artificiale nu vor reuși să construiască o inteligență artificială generală.

În al doilea rând, *Homo sapiens* este curios de la natură, calitate ce va motiva obiectivul științific de a înțelege inteligența și de a dezvolta o inteligență artificială generală chiar și fără stimulente economice. Deși curiozitatea este una dintre cele mai slăvite calități omenești, poate cauza probleme când implică tehnologii pe care nu am învățat cum să le stăpânim înțelept. Curiozitatea științifică pură fără motivația profitului a contribuit la descoperirea armelor nucleare și instrumentelor de inginerie a virusurilor, deci nu este de neconceput ca vechea zicală „curiozitatea a omorât pisica" să se aplice și speciei umane, în cele din urmă.

În al treilea rând, suntem muritori. Acest lucru explică suportul aproape unanim pentru dezvoltarea noilor tehnologii care ne ajută să trăim vieți mai lungi și mai sănătoase, ceea ce motivează puternic cercetările prezente din domeniul inteligenței artificiale. Inteligența artificială generală poate ajuta cercetarea medicală chiar și mai mult. Anumiți gânditori chiar speră să obțină cvasinemurirea prin cyborgizare sau încărcare.

Astfel ne aflăm pe o pantă alunecoasă către inteligența artificială generală, cu stimulente puternice pentru a continua să alunecăm în jos, deși consecința va fi, prin definiție, depășirea noastră economică. Nu va mai fi nevoie de noi pentru nimic, pentru că toate meseriile vor putea fi făcute mai bine de mașini. Crearea reușită a inteligenței artificiale generale ar fi cel mai mare eveniment din istoria umanității, așadar de ce există atât de puține discuții serioase cu privire la unde ne va conduce acest lucru?

Și aici răspunsul implică motive multiple.

În primul rând, cum remarca ironic Upton Sinclair, „este dificil să-l faci pe om să înțeleagă ceva când salariul său depinde de a nu înțelege"*. De exemplu, purtătorii de cuvânt ai companiilor de tehnologie sau ai grupurilor de cercetare universitară pretind adesea că nu există riscuri în activitățile lor chiar dacă în particular gândesc altfel. Observația lui Sinclair poate ajuta la explicarea nu doar a reacțiilor la riscuri determinate de fumat și încălzirea globală, ci și la motivele pentru care unii tratează tehnologia ca pe o nouă religie ale cărei credințe centrale sunt că mai multă tehnologie este întotdeauna mai bine și ai cărei eretici sunt luddiți temători și care nu o înțeleg.

În al doilea rând, oamenii au o istorie veche cu autoiluzionarea, extrapolarea incorectă a trecutului și subestimarea tehnologiilor emergente. Evoluția darwinistă ne-a dotat cu o frică puternică de amenințări concrete, nu de amenințările abstracte ale tehnologiilor viitoare pe care ne este greu să le vizualizăm sau chiar să ni le imaginăm. Gândește-te cum ar fi fost să încerci să-i avertizezi pe oamenii din anii 1930 de o cursă a înarmărilor nucleare viitoare, când nu le-ai fi putut arăta un singur clip video cu o explozie nucleară și nimeni nu știa măcar cum se fabrică astfel de arme. Chiar și cercetătorii de top pot subestima nesiguranța, făcând prognoze care sunt fie prea optimiste – unde sunt reactoarele de fuziune și mașinile zburătoare? –, fie prea pesimiste. Ernest Rutherford, probabil cel mai mare fizician nuclear al vremii sale, spunea în 1933 – cu mai puțin de douăzeci și patru de ore înainte ca Leo Szilard să conceapă reacția nucleară în lanț – că energia nucleară este „o prostie". Efectiv, nimeni din vremea aceea nu a putut anticipa cursa înarmărilor nucleare.

În al treilea rând, psihologii au descoperit că avem tendința să evităm să ne gândim la amenințări neliniștitoare când credem că nu putem face nimic pentru a le evita. Însă, în acest caz, sunt

* Upton Sinclair, *I, Candidate for Governor: And How I Got Licked* (Berkeley, CA: University of California Press, 1994), 109 (n.a.).

multe lucruri constructive pe care le putem face, dacă am putea să începem să ne gândim serios la problemă.

Ce putem face?

Susțin schimbarea strategiei de la „hai să ne grăbim să construim tehnologii care să ne facă să devenim depășiți – ce se poate întâmpla?" la „să ne imaginăm un viitor luminos și să facem eforturi să-l transformăm în realitate".

Pentru a motiva efortul necesar acestei transformări, strategia începe prin imaginarea unei destinații ispititoare. Deși viitorurile portretizate de Hollywood au tendința să fie distopice, adevărul este că inteligența artificială generală poate ajuta viața să înflorească mai mult decât oricând. Tot ce iubesc eu la civilizație este produsul inteligenței, deci, dacă putem să ne amplificăm inteligența prin inteligență artificială generală, avem posibilitatea de a rezolva cele mai spinoase probleme de astăzi și de mâine, inclusiv bolile, încălzirea globală și sărăcia. Cu cât vom face mai detaliată viziunea noastră comună și pozitivă despre viitor, cu atât vom fi mai motivați să lucrăm împreună pentru a o transforma în realitate.

Ce ar trebui să facem pentru a ghida această viziune? Cele douăzeci și trei de principii Asilomar adoptate în 2017 oferă suficientă ghidare, inclusiv următoarele obiective pe termen scurt:

1. O cursă a înarmărilor cu arme autonome letale trebuie evitată.
2. Prosperitatea economică creată de inteligența artificială trebuie distribuită la scară largă, astfel încât întreaga umanitate să se bucure de ea.
3. Investițiile în inteligența artificială trebuie însoțite de fonduri de cercetare speciale destinate asigurării utilizării ei benefice... Cum putem face sistemele viitoare de inteligență artificială foarte robuste, astfel încât să poată face ce vrem noi fără să dea erori de funcționare sau să fie atacate de hackeri?*

* https://futureoflife.org/ai-principles/ (n.a.).

Primele două implică să nu rămânem blocați în echilibre Nash suboptime. O cursă a înarmărilor cu arme autonome letale scăpată de sub control, care determină scăderea continuă a prețurilor asasinatelor anonime automate până la zero, va fi foarte greu de oprit dacă prinde inerție. Al doilea scop ar necesita inversarea tendinței moderne în câteva țări occidentale în care sectoare întregi de populație sunt din ce în ce mai sărace în termeni de sărăcie absolută, alimentând astfel furia, resentimentele și polarizarea. Dacă al treilea obiectiv nu este realizat, toată tehnologia minunată a inteligenței artificiale pe care o creăm ne-ar putea vătăma, fie accidental, fie intenționat.

Cercetarea siguranței inteligenței artificiale trebuie făcută cu un termen-limită strict în minte: înainte de sosirea inteligenței artificiale, trebuie să stabilim cum o vom face să înțeleagă, adopte și păstreze obiectivele noastre. Cu cât mașinile devin mai puternice și mai inteligente, cu atât devine mai importantă alinierea obiectivelor lor cu ale noastre. Cât timp construim mașini relativ lipsite de inteligență, întrebarea nu este dacă obiectivele omenești vor prevala, ci câte probleme ne-ar putea cauza mașinile înainte de a putea rezolva problema alinierii obiectivelor. Însă, dacă o superinteligență este vreodată eliberată, va fi invers: deoarece inteligența este capacitatea de a realiza obiective, o inteligență artificială superinteligentă este prin definiție mult mai bună în realizarea obiectivelor sale decât sunt oamenii la realizarea obiectivelor lor și, prin urmare, va triumfa.

În alte cuvinte, *riscul real al inteligenței artificiale generale nu este rea-voința, ci competența.* O inteligență artificială generală superinteligentă va fi extrem de bună la realizarea obiectivelor sale și, dacă acele obiective nu sunt aliniate cu ale noastre, am încurcat-o. Oamenii nu se gândesc prea mult când inundă mușuroaiele pentru a construi diguri hidroelectrice, așa că am face bine să nu punem umanitatea în locul acelor furnici. Majoritatea cercetătorilor argumentează că, dacă vom crea până la urmă o superinteligență, ar trebui să ne asigurăm că este ceea ce pionierul în domeniul siguranței inteligenței artificiale Eliezer Yudkowsky

a numit „inteligență artificială prietenoasă" – o inteligență artificială ale cărei obiective sunt benefice într-un sens profund.

Problema morală legată de ce ar trebui să fie aceste obiective este la fel de urgentă ca problemele tehnice legate de alinierea obiectivelor. De exemplu, ce fel de societate sperăm să creăm, unde găsim sens și scop în viețile noastre chiar dacă noi, la drept vorbind, nu mai suntem necesari? Adesea primesc următorul răspuns ușuratic la această întrebare: „Să construim mașini mai inteligente decât noi și să le lăsăm pe ele să găsească răspunsul!" Acest răspuns pune greșit semnul de egalitate între inteligență și moralitate. Inteligența nu este bună sau rea, ci neutră din punct de vedere moral. Este pur și simplu capacitatea de a realiza obiective complexe, bune sau rele. Nu putem trage concluzia că lucrurile ar fi stat mai bine dacă Hitler ar fi fost mai inteligent. Într-adevăr, amânarea abordării problemelor de etică până după construirea inteligenței artificiale generale cu obiective aliniate este iresponsabilă și potențial dezastruoasă. O superinteligență perfect obedientă ale cărei obiective se aliniază automat cu cele ale proprietarilor umani ar fi un fel de SS-Obersturmbannführer Adolf Eichmann nazist pe steroizi. În lipsa unei busole morale sau a unor inhibiții personale, ar implementa cu o eficiență nemiloasă obiectivele proprietarilor săi, oricare ar fi acestea.*

Când vorbesc despre nevoia de a analiza riscurile tehnologiei, sunt acuzat că încerc să stârnesc teama. Dar aici, la MIT, unde lucrez, știm că o astfel de analiză a riscurilor nu este o încercare de a stârni teama: este *ingineria siguranței*. Înainte de misiunea de aterizare pe Lună, NASA s-a gândit sistematic la tot ce ar putea să nu meargă când vor pune astronauți în vârful unei rachete de 110 metri, plină cu combustibil foarte inflamabil, și-i va lansa către un loc unde nimeni nu-i putea ajuta – și erau multe lucruri care puteau să nu meargă. Era stârnire a fricii? Nu, era ingineria siguranței care a asigurat succesul misiunii. Similar, ar trebui să

* Vezi, de exemplu, Hannah Arendt, *Eichmann in Jerusalem: A Report on the Banality of Evil* (New York: Penguin Classics, 2006) (n.a.).

analizăm ce ar putea să nu meargă bine cu inteligența artificială pentru a ne asigura că va merge bine.

Perspectivă

Pe scurt, dacă tehnologia depășește înțelepciunea cu care o putem gestiona, ne poate conduce la extincție. Deja a determinat extincția a 20 până la 50 la sută din totalitatea speciilor de pe Pământ, după unele estimări*, și ar fi o ironie a sorții dacă noi am fi următorii. În același timp, ar fi jalnic, dat fiind faptul că oportunitățile oferite de inteligența artificială generală sunt practic astronomice, oferind potențial vieții să prospere miliarde de ani nu doar pe Pământ, ci și într-o mare parte din cosmosul nostru.

În loc să irosim această oportunitate printr-o negare neștiințifică a riscurilor și printr-o planificare proastă, haideți să fim ambițioși! *Homo sapiens* este exaltant de ambițios, cum spun și faimoasele versuri ale lui William Ernest Henley din „Invictus": „Eu, azi, al sorții mele sunt stăpânul, / Sunt căpitanul sufletului meu"**. În loc să ne lăsăm duși de val ca o corabie fără cârmă către propria noastră depășire, să abordăm și să depășim provocările tehnice și sociale care stau între noi și un viitor tehnologizat bun. Dar cum rămâne cu provocările existențiale legate de moralitate, obiective și sens? Nu există niciun sens codat în legile fizicii, așa că, în loc să așteptăm pasiv ca Universul să ne confere nouă sens, să acceptăm și să sărbătorim faptul că noi, ființele conștiente, suntem cele care dăm sens Universului. Să ne creăm propriul sens, bazat pe ceva mai profund decât a avea pur și simplu o meserie. Inteligența artificială generală ne poate permite să devenim, în sfârșit, stăpânii propriului nostru destin. Să facem acest destin unul cu adevărat înălțător!

---

* Vezi Elizabeth Kolbert, *The Sixth Extinction: An Unnatural History* (New York: Henry Holt, 2014) (n.a.).

** În original, „I am the master of my fate, / I am the captain of my soul", traducere de Neculae Leonard (n.t.)

Capitolul 9
Mesaje disidente

Jaan Tallinn

Jaan Tallinn, programator, fizician teoretician și investitor, este cofondator Skype și Kazaa.

Jaan Tallinn a crescut în Estonia, devenind unul dintre puținii dezvoltatori de jocuri pe calculator de acolo, când acea națiune era încă o republică socialistă sovietică. Aici, compară disidenții care au făcut să cadă Cortina de Fier cu disidenții care dau semnalul de alarmă pentru progresele rapide ale inteligenței artificiale. Plasează rădăcinile disidenței curente din domeniul inteligenței artificiale, paradoxal, în rândul unor pionieri ai domeniului inteligenței artificiale precum Wiener, Alan Turing și I.J. Good.

Preocuparea lui Jaan este cu riscurile existențiale, inteligența artificială fiind printre cele mai extreme din mai multe. În 2012, a coînființat Centrul pentru Studierea Riscurilor Existențiale – un institut de cercetare interdisciplinară care lucrează la ameliorarea riscurilor „asociate tehnologiilor emergente și activității umane" – la Universitatea Cambridge, împreună cu filosoful Huw Price și Martin Rees, astronom regal.

S-a descris ca fiind „un consecințialist convins" – suficient de convins încât să-și dea o mare parte din bogăția sa antreprenorială Institutului Future of Life (al cărui cofondator este), Machine Intelligence Research Institute și altor astfel de organizații care lucrează la reducerea riscurilor. Max Tegmark a scris despre el: „Dacă ești o formă de viață inteligentă care citește acest text peste câteva milioane de ani și ești uimit de cum prosperă viața, probabil îți datorezi existența lui Jaan".

Într-o vizită recentă la Londra, eu și Jaan am participat la o conferință pe tema inteligenței artificiale, pentru maratonul Serpentine Gallery organizat de Primăria Londrei, sub egida lui Hans Ulrich Obrist (alt participant la acest volum). Fiind vorba despre lumea artistică, a fost o cină opulentă într-o vilă plină cu oamenii frumoși ai Londrei: artiști, fotomodele, oligarhi, vedete de teatru și de cinema. După ce s-a prezentat în felul său nepretențios („Salut, sunt Jaan"), a spus dintr-odată, „este timpul să dansăm pe hip-hop", s-a aruncat pe podea într-o mână și a început să-și etaleze mișcările spectaculoase în fața personalităților năucite din încăpere. Apoi a mers în subcultura cluburilor de dans, aparent acesta fiind modul în care își încheie fiecare seară când este în turneu. Cine ar fi știut?

În martie 2009, eram într-un restaurant fad de pe marginea unei autostrăzi gălăgioase din California. Eram acolo pentru a mă întâlni cu un tânăr al cărui blog îl urmăream. Pentru a se face recunoscut, purta un nasture pe care scria: *spune adevărul chiar dacă îți tremură vocea.* Numele său era Eliezer Yudkowsky și ne-am petrecut următoarele patru ore discutând despre mesajul pe care îl are el pentru lume – mesajul care mă adusese în acel restaurant și urma să domine activitatea mea ulterioară.

Primul mesaj: ocupația sovietică

În *The Human Use of Human Beings*, Norbert Wiener a privit lumea prin lentila comunicației. A văzut un Univers care mergea în pași de dans pe ritmul celei de-a doua legi a termodinamicii către inevitabila sa moarte termică. Într-un astfel de Univers, singurele entități (meta)stabile sunt *mesajele* – tipare de informație ce se propagă prin timp, ca undele care se propagă pe suprafața unui lac. Chiar și noi, oamenii, putem fi considerați mesaje, deoarece atomii din corpurile noastre sunt prea trecători pentru a ne atașa identitățile noastre lor. În schimb, suntem „mesajul" pe care îl întreține funcțiile noastre corporale. După cum spune Wiener: „tiparul menținut de această homeostază este piatra de hotar a identității noastre personale".

Sunt obișnuit să tratez procesele și calculele ca pe niște elemente fundamentale ale lumii. Acestea fiind spuse, lentila lui

Wiener scoate în evidență câteva aspecte interesante ale lumii care altfel ar fi rămas probabil în fundal și care mi-au conturat viața, într-o mare măsură. Acestea sunt două mesaje, ambele avându-și rădăcinile în cel de-al Doilea Război Mondial. Au început ca fiind mesaje disidente liniștite – mesaje cărora omenii nu le acordau prea mare atenție, chiar dacă erau de acord cu ele tacit sau poate subconștient. Primul mesaj era: *Uniunea Sovietică este compusă dintr-o serie de ocupații ilegitime. Aceste ocupații trebuie să înceteze.*

Eston fiind, am crescut în spatele Cortinei de Fier și am fost în rândul întâi când a căzut. Am auzit acest prim mesaj în reminiscențele nostalgice ale bunicilor mei și între zgomotele aspre care bruiau Vocea Americii. Au fost și mai puternice în timpul erei Gorbaciov, când statul a devenit mai blând în tratamentul aplicat disidenților, crescând apoi în timpul Revoluției Cântate din Estonia, la sfârșitul anilor 1980.

În adolescența mea, am fost martor la răspândirea acestui mesaj în cercuri din ce în ce mai mari de oameni, începând cu disidenții activi, care îl vociferau de jumătate de secol și acest lucru îi costa mult, continuând cu artiștii și oamenii de litere și sfârșind în rândurile membrilor de partid și politicienilor care au schimbat tabăra. Această nouă elită consta într-un grup eclectic de oameni: disidenții inițiali care au reușit să supraviețuiască represiunii, intelectualii publici și (spre exasperarea disidenților supraviețuitori) chiar foști comuniști. Dogmaticii rămași – chiar și cei remarcabili – au fost marginalizați, în cele din urmă, câțiva dintre ei retrăgându-se în Rusia.

Interesant, pe măsură ce mesajul se propaga de la un grup la altul, evolua. A început într-o formă pură și intransigentă („ocupația trebuie să ia sfârșit!") în rândurile disidenților care considerau că adevărul este mai important decât libertatea lor personală. Grupurile populare, care aveau mai mult de pierdut, au calificat și au diluat inițial mesajul, luând poziții ca „ar fi bine pe termen lung să delegăm controlul asupra problemelor locale" (Apăreau întotdeauna excepții: unii intelectuali publici proclamau

mesajul disident inițial *ad litteram*). În cele din urmă, mesajul original – fiind pur și simplu adevărat – a prevalat asupra versiunilor sale diluate. Estonia și-a recâștigat independența în 1991 și ultimele trupe sovietice au plecat după trei ani.

Oamenii care și-au asumat riscuri și au grăit adevărul în Estonia și în alte țări din blocul sovietic au jucat un rol monumental în rezultatul final – un rezultat care a schimbat viețile a sute de milioane de oameni, inclusiv a mea. Au grăit adevărul, chiar dacă le tremurau vocile.

Al doilea mesaj: riscul inteligenței artificiale

Expunerea mea la cel de-al doilea mesaj revoluționar a fost prin intermediul blogului lui Yudkowsky – blogul care m-a făcut să-l contactez și să mă întâlnesc cu el în California. Mesajul era: *Progresul continuu în domeniul inteligenței artificiale poate precipita o schimbare de proporții cosmice – un proces scăpat de sub control care prezintă posibilitatea de a omorî pe toată lumea. Trebuie să facem foarte mult efort suplimentar pentru a evita acest rezultat.*

După întâlnirea mea cu Yudkowsky, primul lucru pe care l-am făcut a fost să încerc să stârnesc interesul colegilor mei de la Skype și colaboratorilor apropiați cu privire la această avertizare. Nu am reușit. Mesajul era prea nebunesc, prea disident. Vremea sa nu venise încă.

Abia mai târziu am aflat că Yudkowsky nu era primul disident care a grăit acest adevăr. În aprilie 2000, în revista *Wired* a apărut un articol de opinie lung intitulat „De ce viitorul nu are nevoie de noi", scris de Bill Joy, cofondatorul și directorul științific al companiei Sun Microsystems. El avertiza:

> „Obișnuiți a trăi cu descoperirile științifice revoluționare, nu ne-am împăcat încă cu faptul că cele mai interesante tehnologii ale secolului al XXI-lea – robotica, ingineria genetică și nanotehnologia – reprezintă o amenințare diferită față de tehnologiile anterioare. În special roboții,

organismele create pe cale inginerească și nanoboții au în comun un factor amplificator periculos: se pot replica automat... un «bot» se poate înmulți și lucrurile pot scăpa rapid de sub control."

Se pare că atacul lui Joy a făcut furori, dar nu a stârnit nicio acțiune.

Însă pentru mine a fost mai surprinzător că mesajul privind riscurile inteligenței artificiale a luat naștere aproape simultan cu domeniul informaticii. Într-o prelegere din 1951, Alan Turing anunța: „Pare probabil ca, odată ce am inițiat gândirea automată, să nu mai dureze mult până ne va depăși puterile noastre plăpânde... Așadar, la un anumit nivel, ar trebui să ne așteptăm ca mașinile să preia controlul..."* După aproximativ un deceniu, colegul său din Bletchley Park, I.J. Good, scria: „prima mașină ultrainteligentă este *ultima* invenție pe care trebuie să o facă omul, dacă mașina este suficient de docilă încât să ne spună cum o putem ține sub control"**. Într-adevăr, am numărat șase locuri în *The Human Use of Human Beings* în care Wiener a făcut aluzie la un aspect sau altul al Problemei Controlului. („Mașina asemănătoare djinnului, care poate învăța și poate lua decizii în baza învățăturii sale, nu va fi obligată în niciun fel să ia deciziile pe care ar fi trebuit să le luăm noi sau care ar fi acceptabile pentru noi.") Aparent, disidenții inițiali care promulgau mesajul despre riscurile inteligenței artificiale sunt chiar pionierii inteligenței artificiale!

Greșeala fatală a evoluției

Au existat multe argumente, unele sofisticate, altele mai puțin, pentru faptul că Problema Controlului este reală, și nu o fantezie

* Retipărită post-mortem în *Philosophia Mathematica* 4, nr. 3 (1966): 256–60 (n.a.).

** Irving John Good, „Speculations Concerning the First Ultraintelligent Machine", Advances in Computers 6 (Cambridge, Massachusetts: Academic Press, 1965): 31–88 (n.a.).

science-fiction. Permiteți-mi să vă prezint una care ilustrează magnitudinea problemei:

În ultima sută de mii de ani, lumea (însemnând Pământul, dar argumentul se poate extinde și la sistemul solar și, probabil, chiar la întregul Univers) a fost în regimul creierului omenesc. În acest regim, creierele ființelor *Homo sapiens* au fost cele mai sofisticate mecanisme de conturare a viitorului (unii le-au numit cele mai complicate obiecte din univers). Inițial, nu le-am folosit la multe altele decât pentru supraviețuire și politică tribală într-un grup de culegători, însă acum efectele lor le depășesc pe cele ale evoluției naturale. Planeta a trecut de la producția de păduri la producția de orașe.

Așa cum a prezis Turing, când vom avea o inteligență artificială supraomenească („metoda gândirii automate"), regimul creierului omenesc se va sfârși. Privește în jurul tău – ești martorul ultimelor decenii dintr-un regim de o sută de mii de ani. Numai acest gând ar trebui să-i facă pe oameni să mediteze înainte de a expedia inteligența artificială doar ca pe un alt instrument. Unul dintre cei mai importanți cercetători din domeniul inteligenței artificiale mi-a mărturisit că s-ar simți foarte ușurat să audă că inteligența artificială de nivel omenesc este imposibil de creat.

Bineînțeles, probabil va mai dura mult timp până vom dezvolta inteligență artificială la nivel omenesc. Dar avem motive să bănuim că nu așa stau lucrurile. La urma urmei, nu a durat mult, în termeni relativi, ca evoluția – procesul de optimizare orb și stângaci – să creeze inteligența omenească după ce a avut animalele ca materie primă. Sau chiar viața pluricelulară: se pare că i-a fost mult mai greu să facă celulele să se alipească decât i-a fost să creeze inteligența omenească după ce a avut organismele pluricelulare. Fără să mai menționăm faptul că nivelul nostru de inteligență era limitat de factori grotești precum lățimea canalului de naștere. Imaginează-ți un dezvoltator al inteligenței artificiale care nu mai poate avansa pentru că nu poate ajusta mărimea fontului pe calculatorul său!

Există aici o simetrie interesantă: în producția de oameni, evoluția a creat un sistem care este, cel puțin în multe dimensiuni importante, un planificator și un organizator mai puternic decât evoluția însăși. Suntem prima specie care a înțeles că este produsul evoluției. În plus, am creat multe artefacte (radiouri, arme de foc, nave spațiale) pe care evoluția avea slabe șanse să le creeze. Prin urmare, viitorul nostru va fi determinat de propriile noastre decizii, și nu de evoluția noastră biologică. În acest sens, evoluția a căzut victimă propriei sale Probleme a Controlului.

Putem doar să sperăm că suntem mai inteligenți decât evoluția în această privință. *Suntem* mai inteligenți, bineînțeles, dar va fi oare suficient? Suntem pe cale de a afla.

Situația curentă

Așadar, iată-ne, la peste jumătate de secol după avertizările inițiale ale lui Turing, Wiener și Good, și la un deceniu după ce oameni ca mine au început să acorde atenție mesajului privind riscurile inteligenței artificiale. Mă bucur să văd că am făcut multe progrese în confruntarea acestor probleme, dar cu siguranță nu le-am dat de cap încă. Riscurile inteligenței artificiale, deși nu mai sunt un subiect tabu, nu sunt încă apreciate la valoarea lor reală în rândurile cercetătorilor din domeniu. Riscurile inteligenței artificiale nu sunt nici lucruri de care să știe toată lumea. În raport cu cronologia primului nostru mesaj disident, aș spune că suntem aproximativ în anul 1988, când pornirea unei discuții despre ocupația sovietică nu mai era ceva ce ți-ar putea distruge cariera, dar trebuia să-ți acoperi totuși cumva poziția. Aud acoperiri similare și în ziua de azi – afirmații precum „nu-mi fac probleme legate de o superinteligență artificială, dar există probleme etice reale determinate de gradul din ce în ce mai mare de automatizare" sau „este bine că unii oameni cercetează și riscurile inteligenței artificiale, dar nu este o îngrijorare pe termen scurt" sau chiar ceva ce sună rezonabil precum „acestea sunt scenarii cu o

probabilitate foarte mică, dar potențialul impact major justifică atenția acordată".

Cât despre propagarea mesajului, ne apropiem de punctul de cotitură. Un studiu recent realizat pe cercetătorii inteligenței artificiale care au publicat la cele două conferințe internaționale majore pe tema inteligenței artificiale din 2015 arată că 40% dintre aceștia consideră acum că riscurile unei inteligențe artificiale superavansate sunt fie „o problemă importantă", fie „una dintre cele mai importante probleme din domeniu"*

Bineînțeles, la fel cum au existat comuniști dogmatici care nu și-au schimbat niciodată poziția, este garantat și faptul că anumiți oameni nu vor accepta niciodată ideea că inteligența artificială poate fi periculoasă. Mulți dintre cei care au negat comunismul veneau din nomenclatura sovietică; similar, cei care neagă riscurile inteligenței artificiale au adesea motive financiare sau alte motive pragmatice. Unul dintre principalele motive este reprezentat de profiturile corporațiilor. Inteligența artificială este profitabilă și, chiar și în situațiile în care nu este, este cel puțin o întreprindere trendy, orientată către viitor, cu care să-ți asociezi compania. Astfel multe dintre pozițiile de negare sunt produse ale departamentului de relații publice din companii și mașinații juridice. Într-un sens foarte real, marile corporații sunt mașini nonumane care-și urmăresc propriile interese – interese care pot să nu fie aliniate cu cele ale oamenilor care lucrează pentru ele. Așa cum observa Wiener în *The Human Use of Human Beings*: „când atomii omenești sunt croșetați într-o organizație în care sunt utilizați, nu în dreptul lor deplin de ființe umane responsabile, ci ca rotițe, tije și manete, contează prea puțin că materia lor primă este din carne și sânge".

Un alt stimulent puternic pentru a închide ochii la riscurile inteligenței artificiale este curiozitatea (foarte omenească) care nu cunoaște nicio limită. „Când vezi ceva care este atrăgător din

* Katja Grace et al., „When Will AI Exceed Human Performance? Evidence from AI Experts", https://arxiv.org/pdf/1705.08807.pdf (n.a.).

punct de vedere tehnic, îl transformi în realitate și începi să discuți despre ce ar trebui făcut doar după ce ai obținut succesul tehnic. La fel au stat lucrurile și cu bomba atomică", a spus J. Robert Oppenheimer. Cuvintele sale au fost repetate recent de Geoffrey Hinton, despre care s-ar putea spune că este inventatorul învățării profunde, în contextul riscurilor inteligenței artificiale: „ți-aș putea da argumentele obișnuite, dar adevărul este că ideea descoperirii este mult prea atrăgătoare".

Incontestabil, putem mulțumi atitudinii antreprenoriale și curiozității științifice pentru aproape toate lucrurile frumoase pe care le luăm de bune în epoca modernă. Însă, este important să ne dăm seama că progresul nu ne *datorează* un viitor bun. În cuvintele lui Wiener, „este posibil să crezi în progres ca stare de fapt fără să crezi în progres ca principiu etic".

În cele din urmă, nu ne permitem luxul de a aștepta până când toți corporatiștii și cercetătorii inteligenței artificiale vor fi de acord să accepte riscurile inteligenței artificiale. Imaginează-ți că stai într-un avion care urmează să decoleze. Dintr-odată, auzi un anunț care spune că 40% dintre experți consideră că la bord se află o bombă. În acel moment, e clar ce ai de făcut și e clar că nu vei sta acolo să aștepți să cadă de acord și ceilalți 60%.

Calibrarea mesajului privind riscurile inteligenței artificiale

Deși este uimitor de profetic, mesajul privind riscurile inteligenței artificiale venit de la disidenții inițiali are un mare defect – la fel ca versiunea care domină discursul public curent: ambele subestimează considerabil magnitudinea problemei, precum și avantajele potențiale ale inteligenței artificiale. Mesajul, în alte cuvinte, nu transmite adecvat adevăratele mize ale jocului.

Wiener a avertizat inițial în legătură cu riscurile *sociale* – riscurile care își vor avea originea în integrarea neglijentă a deciziilor generate de mașini cu procesele de guvernare și proasta utilizare (de către oameni) a acestui proces automat de luare a deciziilor.

Asemenea, dezbaterea „serioasă" curentă cu privire la riscurile inteligenței artificiale se concentrează în primul rând pe lucruri precum șomaj tehnologic sau prejudecăți ale învățării automate. Deși astfel de discuții pot fi valoroase și pot aborda probleme serioase pe termen scurt, sunt extrem de înguste. Îmi aduc aminte de remarca lui Yudkowsky dintr-o postare pe blogul său: „să te întrebi care sunt efectele *superinteligenței automate* asupra unei piețe de muncă omenești convenționale este ca și cum te-ai întreba cum vor fi afectate tiparele de comerț dintre SUA și China dacă s-ar izbi Luna de Pământ. Cu siguranță ar exista efecte, dar nu ai înțeles ideea principală".

Din punctul meu de vedere, punctul central al riscurilor inteligenței artificiale este acela că *inteligența artificială superinteligentă este un risc de mediu.* Permiteți-mi să explic.

În parabola sa despre „balta conștientă", Douglas Adams descrie o baltă ce se trezește într-o dimineață într-o groapă care i se potrivește „surprinzător de bine". Din acea observație, balta trage concluzia că lumea trebuie să fie făcută pentru ea. Prin urmare, scrie Adams, „momentul în care dispare o ia pe nepregătite". Să presupunem că riscurile inteligenței artificiale sunt limitate doar la dezvoltările sociale adverse înseamnă să facem o greșeală similară. Realitatea crudă este că universul nu este făcut pentru noi; în schimb, suntem ajustați de evoluție pentru a supraviețui într-un interval foarte precis de parametri de mediu. De exemplu, avem nevoie ca atmosfera de la sol să fie aproximativ la temperatura camerei, să aibă aproximativ 100 kPa presiune și o concentrație suficientă de oxigen. Orice schimbare, chiar și temporară, a acestui echilibru precar și murim în câteva minute.

Inteligența bazată pe siliciu nu împărtășește aceleași probleme legate de mediu. Acesta este motivul pentru care este mult mai ieftin să explorăm spațiul cu sonde automate, și nu cu „conserve de carne". În plus, mediul actual al Pământului este aproape sigur suboptim pentru ceea ce ar vrea o inteligență artificială superinteligentă: *putere de calcul.* Așadar, ne putem trezi că planeta noastră

trece de la o încălzire globală antropogenetică la o răcire globală mașinogenetică. O provocare importantă pe care trebuie să o abordeze specialiștii în siguranța inteligenței artificiale este cum ar putea să țină sub control o potențială inteligență artificială superinteligentă – o inteligență artificială cu o amprentă mult mai mare decât a noastră –, astfel încât să nu transforme mediul nostru într-o zonă care nu mai poate fi locuită de forme de viață biologice.

În mod interesant, dat fiind faptul că cele mai puternice surse atât pentru cercetarea inteligenței artificiale, cât și pentru negarea riscurilor sale, sunt sub marile umbrele corporative, dacă-ți încordezi privirea suficient, mesajul „inteligenței artificiale ca risc de mediu" arată ca îngrijorarea cronică legată de corporațiile care se eschivează în fața responsabilității lor față de mediu.

În sensul invers, grijile despre efectele sociale ale inteligenței artificiale nu percep nici părțile pozitive în totalitate. Este greu să exagerezi cât de mic și de local este viitorul planetei noastre, comparat cu potențialul deplin al umanității. La scări temporale astronomice, planeta noastră va dispărea în curând (dacă nu cumva învățăm să îmblânzim Soarele, lucru care este și el posibil) și aproape toate resursele – atomi și energie liberă – care vor susține civilizația pe termen lung sunt în spațiul adânc.

Eric Drexler, inventatorul nanotehnologiei, a popularizat recent conceptul de „pareto-topie": ideea conform căreia inteligența artificială, dacă este făcută bine, poate determina un viitor în care viețile *tuturor* oamenilor sunt mult îmbunătățite, un viitor în care nu mai există pierzători. O realizare-cheie în această privință este că principalul motiv pentru care umanitatea nu-și atinge potențialul deplin poate fi simțul nostru instinctiv că suntem într-un joc cu sumă zero – un joc în care jucătorii trebuie să obțină mici victorii unii pe seama altora. Un astfel de instinct este foarte dezorientat și distructiv într-un „joc" unde miza este totul, iar câștigul potențial este astronomic. Sunt mult mai multe sisteme solare în galaxia noastră decât sunt oameni pe Pământ.

Speranța

Ca și în această scriere, sunt un optimist prevăzător cu privire la faptul că mesajul riscurilor inteligenței artificiale poate salva umanitatea de la extincție, la fel cum mesajul ocupației sovietice a sfârșit prin a elibera sute de milioane de oameni. Începând cu 2015, a atins și a convertit 40% dintre cercetătorii inteligenței artificiale. Nu m-ar surprinde dacă un nou studiu ar arăta că majoritatea cercetătorilor din domeniul inteligenței artificiale consideră că siguranța inteligenței artificiale este o problemă importantă.

Sunt încântat să văd că urmează să apară primele lucrări tehnice pe tema siguranței inteligenței artificiale de pe urma DeepMind, OpenAI și Google Brain și a spiritului colaborator de soluționare a problemelor, spirit care înflorește în echipele de cercetare a siguranței inteligenței artificiale din aceste organizații care altfel ar fi foarte competitive.

Și elita politică și comercială a lumii se trezește încet-încet: siguranța inteligenței artificiale a fost acoperită în rapoarte și prezentări realizate de Institute of Electrical and Electronics Engineers (IEEE), World Economic Forum și de Organization for Economic Cooperation and Development (OECD). Chiar și recentul (iulie 2017) manifest chinez privind inteligența artificială conține secțiuni speciale privind „supravegherea siguranței inteligenței artificiale", „dezvoltarea unor legi, reglementări și norme etice" și stabilirea unui „sistem de evaluare și securitate a inteligenței artificiale" pentru, printre altele, a „spori conștientizarea riscurilor". Sper din inimă ca noua generație de lideri care înțeleg problema controlului inteligenței artificiale și inteligența artificială ca risc ultim de mediu să depășească obișnuitele jocuri tribale, cu sumă zero, și să direcționeze umanitatea dincolo de aceste ape periculoase în care ne aflăm – deschizându-ne astfel calea către stelele care ne-au așteptat miliarde de ani.

Un toast pentru următoarea noastră sută de mii de ani! Și nu ezita să grăiești adevărul, chiar dacă-ți tremură vocea.

Capitolul 10
Profeția tehnologică și puterea cauzală subapreciată a ideilor

Steven Pinker

Steven Pinker, profesor Johnstone Family în Departamentul de Psihologie de la Universitatea Harvard, este un psiholog experimental, care face cercetări în domeniul cogniției vizuale, psiholingvisticii și relațiilor sociale. Este autorul a unsprezece cărți, printre care *The Blank Slate, The Better Angels of Our Nature* și, recent, *Enlightenment Now: The Case for Reason, Science, Humanism, and Progress.*

În decursul carierei sale, studiind limbajul, susținând o biologie realistă a minții sau examinând condiția umană prin lentila ideilor umaniste ale iluminismului, psihologul Steven Pinker a îmbrățișat și a promovat o înțelegere naturalistă a universului și o teorie a minții computațională. Este probabil primul intelectual public recunoscut internațional a cărui recunoaștere este bazată pe sprijinirea gândirii bazate empiric despre limbaj, minte și natura umană.

„La fel cum Darwin a făcut posibil ca un observator atent al lumii să se descurce și fără creaționism", spune el, „Turing și alții au făcut posibil ca un observator atent al

lumii cognitive să se descurce fără spiritualitate". În dezbaterea despre riscurile inteligenței artificiale, vorbește împotriva profețiilor apocaliptice, observând că izvorăsc din cele mai întunecate înclinații psihologice ale noastre – exemplificând în principiu prin rapoartele mass-media: „scenariile dezastruoase sunt ieftine în zona fără probabilități a imaginației noastre și pot găsi întotdeauna o audiență îngrijorată, tehnofobă sau fascinată morbid". De aici au apărut de-a lungul secolelor: Pandora, Faust, Ucenicul Vrăjitor, Frankenstein, bomba populației, epuizarea resurselor, HAL, valizele cu focoase nucleare, teama de anul 2000 și de sufocarea într-o mâzgă nanotehnologică gri. „O caracteristică a distopiilor inteligenței artificiale", evidențiază el „este că proiectează o psihologie îngustă de mascul alfa în conceptul de inteligență... Istoria mai scoate la iveală din când în când câte un despot megaloman sau câte un criminal în serie psihopat, dar acestea sunt produsele unei istorii a selecției naturale care dă formă unor circuite sensibile la testosteron într-o anumită specie de primate, nu unei caracteristici inevitabile a sistemelor inteligente".

În eseul de față, aplaudă credința lui Wiener în forța ideilor vizavi de tehnologia uzurpatoare. În bine alesele cuvinte ale lui Wiener, „Pericolul reprezentat de mașini pentru societate nu vine de la mașina în sine, ci de la ce face omul cu mașina".

Inteligența artificială este dovada existenței uneia dintre ideile măreț ale istoriei umane: ideea că tărâmul abstract al cunoștinței, rațiunii și scopului nu constă în elanul vital, sufletul imaterial sau puterile miraculoase ale țesutului neuronal. În schimb, poate fi legat de tărâmul fizic al animalelor și mașinilor prin concepte precum informația, calculul și controlul. Cunoașterea poate fi explicată ca o serie de tipare în materie sau energie care au o relație sistemică în raport cu stările lumii, cu adevăruri matematice și logice sau unul cu altul. Raționarea poate fi explicată ca transformări ale cunoștinței prin operații fizice concepute să păstreze aceste relații. Scopul poate fi explicat drept controlul operațiilor de a implementa schimbări în lume, ghidat de discrepanțele dintre starea curentă și starea obiectivului. Creierele evoluate natural sunt pur și simplu cele mai familiare sisteme care obțin inteligență prin informație, calcul și control. Sistemele concepute de oameni ce obțin inteligența exonerează ideea că procesarea informațiilor este suficientă pentru a o explica – idee pe care răposatul Jerry Fodor a denumit-o teoria computațională a minții.

Piatra de hotar a acestui volum, cartea lui Norbert Wiener *The Human Use of Human Beings*, a sărbătorit această realizare intelectuală, al cărei contribuitor principal a fost Wiener însuși. O istorie simplă a evoluției din secolul XX care a dat lumii teoria computațională a minții ar putea recunoaște teoriile lui Claude Shannon și Warren Weaver în explicarea cunoștinței și comunicării în termeni de informație. Ar putea recunoaște teoriile lui Alan

Turing și John von Neumann care explică inteligența și puterea de judecată în termeni computaționali. Și ar trebui să recunoască teoria lui Wiener care explică până în prezent lumea misterioasă a scopurilor, obiectivelor și teleologiei în funcție de conceptele tehnice de *feedback*, control și cibernetică (în sensul său original de „guvernare" a funcționării unui sistem direcționat de un obiectiv). „Teoria mea", a anunțat el, „este că funcționarea fizică a individului viu și funcționarea unora dintre mașinile de comunicare mai noi sunt perfect paralele în încercările lor analoge de a controla entropia prin *feedback*" – prevenirea entropiei secătuitoare de viață fiind obiectivul ultim al ființelor umane.

Wiener a aplicat ideea ciberneticii unui sistem terț: societatea. Legile, normele, obiceiurile, mass-media, forumurile și instituțiile unei comunități complexe pot fi considerate canale de propagare a informațiilor și a *feedback*-ului care permit unei societăți să evite dezordinea și să urmărească anumite obiective. Acesta este firul roșu care unește ideile cărții și pe care Wiener l-a văzut ca fiind principala sa contribuție. Explicând *feedback*-ul, scria: „acest complex al comportamentului este ignorat de omul obișnuit și nu joacă rolul pe care ar trebui să-l joace în analiza noastră obișnuită a societății; deoarece, la fel cum răspunsurile fizice individuale pot fi privite din acest punct de vedere, la fel pot fi privite și răspunsurile societății".

Într-adevăr, Wiener conferă greutate ideii conform căreia, în domeniile istoriei, politicii și societății, *ideile contează*. Credințele, ideologiile, normele, legile și obiceiurile, modificând comportamentul oamenilor care le adoptă, pot da contur unei societăți și pot alimenta cursul evenimentelor istorice la fel de sigur cum fenomenele fizice afectează structura și evoluția sistemului solar. A spune că ideile – și nu doar vremea, resursele, geografia sau armele – pot afecta istoria nu este un misticism ieftin. Este o afirmație a puterilor cauzale ale informațiilor instalate în creierele omenești și schimbate în rețele de comunicare și *feedback*. Teoriile deterministe asupra istoriei, indiferent dacă identifică motorul cauzal ca tehnologic, climateric sau geografic, sunt contrazise de puterea cauzală

a ideilor. Efectele acestor idei pot include răsturnări și oscilații impredictibile care pot lua naștere din *feedback*-ul pozitiv sau din *feedback*-ul negativ incorect calibrat.

O analiză a societății în funcție de propagarea ideilor i-a conferit lui Wiener și o indicație pentru critica socială. O societate sănătoasă – una care conferă membrilor săi șansa de a-și trăi viața sfidând entropia – permite informațiilor intuite și aduse de membrii săi să se întoarcă prin *feedback* și să afecteze cumva societatea. O societate disfuncțională invocă dogma și autoritatea pentru a impune controlul de sus în jos. Astfel Wiener s-a descris drept „participant într-o perspectivă liberală" și și-a dedicat cea mai mare parte a energiei sale morale și retorice din carte (atât în ediția din 1950, cât și în cea din 1954) denunțării comunismului, fascismului, mccarthyismului, militarismului și religiei autoritare (în special catolicismului și islamului), avertizând că instituțiile politice și științifice deveneau prea ierarhice și „insulare".

Pe alocuri, cartea lui Wiener mai este și un exemplar timpuriu al unui gen din ce în ce mai popular, profețiile tehnologice. Profeție nu în sensul de simplă prognoză, ci în sensul din Vechiul Testament de avertizare sumbră a unei retribuții catastrofice pentru decadența contemporanilor. Wiener ne-a avertizat de pericolele cursei accelerate de înarmare nucleară, de modificările tehnologice care s-au impus fără a lua în considerare bunăstarea omului („Trebuie să știm, în calitatea noastră de oameni de știință, care este natura umană și care sunt scopurile sale inerente") și de ceea ce astăzi este numit problema alinierii valorilor: că „mașina asemenea djinnului, care poate învăța și poate lua decizii în baza învățării sale, nu va fi obligată în niciun fel să ia deciziile pe care ar fi trebuit să le luăm noi sau ar fi acceptabile pentru noi". În ediția mai sumbră din 1950, ne-a avertizat de pericolul unui „fascism nou și amenințător dependent de *machine à gouverner*".

Profețiile tehnologice ale lui Wiener își au rădăcinile în rebeliunea mișcării romantice împotriva „morilor satanice întunecate" ale Revoluției Industriale, și poate chiar și mai devreme, în arhetipurile lui Prometeu, Pandora și Faust. Iar astăzi au prins viteză.

Câțiva Ieremia, mulți dintre ei (ca și Wiener) din lumea științei și a tehnologiei, au tras semnale de alarmă cu privire la nanotehnologie, inginerie genetică, Big Data și în special la inteligența artificială. Mai mulți participanți la acest volum caracterizează cartea lui Wiener ca fiind un exemplu vizionar de profeție tehnologică și amplifică îngrijorările sale întunecate.

Însă cele două teme morale din *The Human Use of Human Beings* – apărarea liberală a unei societăți deschise și teama distopică de tehnologie scăpată de sub control – sunt în conflict. O societate cu canale de *feedback* care maximizează prosperitatea omenească va avea implementate mecanisme și le va putea adapta unor circumstanțe schimbătoare într-un mod care poate domestici tehnologia, adaptând-o nevoilor omenești. Și aceasta nu este o speranță idealistă sau mistică; cum spunea Wiener, ideile, normele și instituțiile sunt ele însele o formă de tehnologie care constă în tipare de informație distribuite în creiere. Posibilitatea ca mașinile să introducă un nou fascism trebuie comparată în opoziție cu vigoarea ideilor, instituțiilor și normelor liberale pe care Wiener le-a susținut în carte. Hiba din profețiile distopice de astăzi este că ignoră existența acestor norme și instituții sau subestimează grav potența lor cauzală. Rezultatul este un determinism tehnologic ale cărui predicții întunecate sunt respinse în mod repetat de cursul evenimentelor. Numerele „1984" și „2001" sunt mementouri bune.

Voi analiza două exemple. Profeții tehnologici avertizează adesea de un „stat supraveghetor" în care guvernele împuternicite de tehnologie vor monitoriza și interpreta toate comunicațiile private, permițându-le să identifice disidenții și subversivii pe măsură ce apar și făcând irelevantă rezistența la puterea statului. Teleecranele lui Orwell sunt prototipul și, în 1976, Joseph Weizenbaum, unul dintre cei mai sumbri profeți tehnologici ai tuturor timpurilor, a avertizat seria mea de absolvenți să nu urmărească recunoașterea automată a vorbirii, deoarece supravegherea guvernamentală era singura aplicație conceptibilă.

Deși sunt oficial un libertarian civil, profund preocupat de amenințările contemporane la adresa exprimării libere, nu mă

înspăimântă avansurile tehnologice ale internetului, tehnologiei video sau inteligenței artificiale. Motivul este că aproape toate variațiile în spațiu și timp ale libertății de gândire au la bază diferențele normelor și instituțiilor, și aproape niciuna diferențele de tehnologie. Deși ne putem imagina combinații ipotetice ale celor mai răuvoitori totalitari cu cele mai avansate tehnologii, în lumea reală trebuie să fim atenți la norme și legi, nu la tehnologie.

Să ne gândim la variațiile în timp. Dacă, așa cum sugera Orwell, avansul tehnologiei ar fi fost un facilitator principal al represiunii politice, atunci societățile occidentale ar fi trebuit să fie din ce în ce mai restrictive cu exprimarea de-a lungul secolelor, cu o înrăutățire gravă a în a doua jumătate a secolului XX și continuând în secolul al XXI-lea. Nu așa s-a petrecut istoria. În secolele în care comunicarea era implementată cu ajutorul penei și cernelii au existat autodafeuri ale gânditorilor iluminiști, aruncarea lor în închisoare sau chiar ghilotinarea lor. În timpul Primului Război Mondial, când tehnologia fără fir era cea mai nouă, Bertrand Russell era încarcerat pentru opiniile sale pacifiste. În anii 1950, când computerele erau mașini de calcul care ocupau o cameră întreagă, sute de scriitori și erudiți liberali erau pedepsiți profesional. Însă, în secolul al XXI-lea, hiperconectat și cu o accelerare tehnologică explozivă, optsprezece procente dintre profesorii de științe sociale sunt marxiști;* președintele Statelor Unite ale Americii este ridiculizat în fiecare seară de comedianții de la televiziuni ca fiind rasist, pervers și idiot; și cea mai mare amenințare a tehnologiei la adresa discursului politic vine din amplificarea a prea multe voci dubioase, și nu din suprimarea celor luminate.

Acum să ne gândim la variațiile în funcție de loc. Țările occidentale care se află la frontiera tehnologică obține consecvent cele mai mari punctaje în ceea ce privește indicii democrației și drepturilor omului, în timp ce multe state conduse de lideri puternici,

* Neil Gross și Solon Simmons, „The Social and Political Views of American College and University Professors", în *Professors and Their Politics*, ed. N. Gross și S. Simmons (Baltimore: Johns Hopkins University Press, 2014) (n.a.).

dar mai înapoiate sunt în partea de jos a clasamentului, omorând și întemnițând tot timpul criticii guvernului. Lipsa unei corelații între tehnologie și represiune nu este deloc surprinzătoare când analizăm fluxul canalelor de informații din orice societate umană. Pentru ca disidenții să fie influenți, trebuie să-și transmită mesajul într-o rețea vastă prin orice mijloc de comunicare este disponibil – pamfletărie, oratorie cu lădița sub picioare, serate subversive în cafenele și pub-uri, mesaje orale. Aceste canale îi prind ca într-o plasă disidenții influenți dintr-o rețea socială întinsă, lucru care îi face ușor de identificat și de urmărit. Cu atât mai mult atunci când dictatorii redescoperă tehnica veche a transformării oamenilor în arme folosite unul împotriva celuilalt, pedepsindu-i pe cei care nu denunță sau nu-i pedepsesc pe alții.

În contrast, societățile avansate din punct de vedere tehnologic au de mult timp mijloacele de a instala camere de supraveghere conectate prin internet și monitorizate de stat în orice bar și orice dormitor. Însă acest lucru nu s-a întâmplat, deoarece guvernele democratice (chiar și administrația americană prezentă cu impulsurile sale flagrant antidemocratice) nu au mijloacele de a forța supravegherea unui popor zgomotos și greu de controlat, obișnuit să spună ce vrea. Ocazional, pericolele terorismului nuclear, biologic sau cibernetic pun presiune asupra agențiilor naționale de securitate să ia măsuri cum ar fi urmărirea metadatelor de telefonie mobilă, însă aceste măsuri ineficiente, mai mult teatrale decât opresive, nu au avut efecte semnificative nici asupra securității, nici asupra libertății. Ironia sorții este că profețiile tehnologice joacă un rol important în încurajarea unor astfel de măsuri. Semănând panică despre așa-zisele amenințări existențiale cum ar fi bombe nucleare în valize și arme biologice asamblate în dormitoarele adolescenților, pun presiune asupra guvernelor să dovedească faptul că fac ceva, orice, pentru a proteja poporul american.

Ideea nu este că libertatea politică are grijă de ea însăși, ci că cele mai mari amenințări rezidă în rețelele de idei, de norme și instituții care permit *feedback*-ul informației reflectat sau nu de deciziile și înțelegerea colective. În contrast cu himera

amenințărilor tehnologice, o amenințare reală în ziua de astăzi este corectitudinea politică opresivă, care a sugrumat gama largă de ipoteze care se pot exprima în public, a înspăimântat multe persoane inteligente până la punctul în care nu mai vor să intre în arena intelectuală și a declanșat o reacție negativă puternică. O altă amenințare reală este combinarea discreției organelor procuraturii cu o lege din ce în ce mai stufoasă, plină de statute vagi. Rezultatul este că fiecare american comite fără să știe „trei infracțiuni pe zi" (conform titlului cărții scrise de libertarianul civil Harvey Silverglate) și este în pericol de a fi băgat la închisoare oricând guvernul are nevoie de acest lucru. Această armă a organelor procuraturii îl face pe Fratele cel Mare atotputernic, și nu teleecranele*. Activismul și polemizarea direcționate împotriva programelor de supraveghere ale guvernelor ar fi mai bine direcționate către puterile sale juridice înfumurate.

Celălalt punct de focalizare al mai multor profeții tehnologice din zilele noastre este inteligența artificială, fie ea cea din distopiile SF originale cu computere care umblă dezlănțuite și ne înrobesc într-o cursă continuă către dominație, fie în noua versiune, în care ne subjugă accidental, urmărind cu ochelari de cal un obiectiv primit de la noi, fără a ține cont de efectele secundare asupra bunăstării umane (problema alinierii valorilor sugerată de Wiener). Și aici ambele amenințări mi se par himerice, crescând dintr-un determinism tehnologic restrâns care neglijează rețelele de informații și control într-un sistem inteligent, cum ar fi un computer sau un creier, și într-o societate ca întreg.

Frica de subjugare se bazează pe o concepție confuză a inteligenței care datorează mai multe Marelui Lanț al Ființei și voinței de putere a lui Nietzsche decât analizei pe care o face Wiener inteligenței și scopului în termeni de informație, calcul și control. În aceste scenarii înspăimântătoare, inteligența este portretizată ca o potențiune atotputernică, îndeplinind dorințe și pe care agenții o

* Termen folosit de Mihnea Gafița în traducerea romanului *1984*, de George Orwell (n.t.).

posedă în cantități diferite. Oamenii au mai mult din ea decât animalele și un computer sau robot creat artificial va avea mai mult din ea decât oamenii. Deoarece noi, oamenii, ne-am folosit zestrea noastră moderată pentru a domestici sau extermina animale mai puțin dotate decât noi (și deoarece societățile avansate din punct de vedere tehnologic au înrobit sau anihilat societățile cu tehnologii primitive), rezultă că o inteligență artificială superinteligentă ne-ar face același lucru nouă. Deoarece inteligența artificială va gândi de milioane de ori mai rapid decât noi și își va folosi superinteligența pentru a-și îmbunătăți recursiv superinteligența, din momentul în care este activată nu o mai putem opri.

Dar aceste scenarii se bazează pe confundarea inteligenței cu motivația – a credințelor cu dorințele, a deducțiilor cu obiectivele, tehnica de calcul elucidată de Turing și controlul elucidat de Wiener. Chiar dacă am inventa roboți cu o inteligență supraomenească, de ce *ar vrea* să-și înrobească stăpânii sau să preia controlul asupra lumii? Inteligența este abilitatea de a implementa mijloace inedite în atingerea unui obiectiv. Dar obiectivele sunt în afara inteligenței: a fi inteligent nu este același lucru cu a vrea ceva. Se întâmplă ca inteligența speciei *Homo sapiens* să fie un produs al selecției naturale darwiniste, un proces inerent competitiv. În creierul acestei specii, facultatea de judecată vine la pachet cu obiective precum dominarea rivalilor și colectarea resurselor. Dar este o greșeală să confunzi un circuit din creierul limbic al unei anumite specii de primate cu esența inteligenței. Nu există nicio lege a sistemelor complexe care spune că agenții inteligenți trebuie să se transforme în megalomani fără scrupule.

O a doua idee greșită este să credem că inteligența este un continuum nelimitat de potență, un elixir miraculos cu puterea de a rezolva orice problemă, de a atinge orice obiectiv. Eroarea conduce la întrebări fără sens, de exemplu, legate de momentul în care „inteligența artificială depășește inteligența omenească" și la imaginea unei „inteligențe artificiale generale" cu omniscientă și omnipotență dumnezeiești. Inteligența este un sistem inventiv de dispozitive: module *software* ce colectează sau sunt programate să

aibă cunoștințe despre cum să urmărească diverse obiective din diverse domenii. Oamenii sunt echipați să găsească mâncare, să câștige prieteni și să influențeze oameni, să-i farmece pe potențialii parteneri, să crească copii, să se deplaseze în lume și să urmărească alte obsesii și plăceri omenești. Computerele pot fi programate să preia unele dintre aceste probleme (cum ar fi recunoașterea fețelor), să nu se deranjeze cu altele (cum ar fi fermecarea partenerilor) și să preia în plus alte probleme pe care oamenii nu le pot aborda (cum ar fi simularea climei sau sortarea a milioane de înregistrări contabile). Problemele sunt diferite și tipurile de cunoștințe necesare pentru a le rezolva sunt diverse.

Dar, în loc să recunoaștem caracterul central al cunoașterii în inteligență, scenariile distopice confundă o inteligență artificială generală viitoare cu demonul lui Laplace, ființa mitică ce cunoaște locul și mișcarea fiecărei particule din univers și le folosește ca termeni ai unor ecuații de legi fizice pentru a calcula starea oricărui lucru în orice moment din viitor. Din multe motive, demonul lui Laplace nu va fi implementat niciodată în siliciu. Un sistem inteligent real trebuie să obțină informații despre lumea dezordonată a obiectelor și oamenilor interacționând cu ea domeniu cu domeniu, ciclul fiind guvernat de ritmul în care se desfășoară evenimentele în lumea fizică. Acesta este unul dintre motivele pentru care înțelegerea nu respectă Legea lui Moore: cunoștințele sunt obținute prin formularea explicațiilor și testarea acestora în realitate, nu prin executarea unui algoritm din ce în ce mai repede. Devorarea informațiilor de pe internet nu va conferi nici ea omnisciență: Big Data sunt totuși date finite, iar universul cunoașterii este infinit.

Un al treilea motiv pentru care ar trebui să fim sceptici la ideea de preluare bruscă a controlului de către inteligența artificială este că ia mult prea în serios faza de inflație din ciclul *hype* al inteligenței artificiale în care trăim astăzi. În ciuda progresului din domeniul învățării automate, mai ales al rețelelor neuronale artificiale cu straturi multiple, sistemele curente de inteligență artificială sunt departe de a atinge o inteligență generală (dacă acest concept este

măcar coerent). În schimb, sunt restricționate la probleme care constau în maparea unor intrări bine definite în ieșiri bine definite în domenii în care sunt disponibile seturi de instruire gargantuești, în care metrica succesului este imediată și precisă, în care mediul nu se schimbă și în care nu este necesar un proces de raționare treptat, ierarhic sau abstract. Multe succese nu au izvorât dintr-o înțelegere mai bună a mecanismelor de funcționare a inteligenței, ci din puterea forței brute a cipurilor mai rapide și a seturilor din ce în ce mai mari de date, care permit programelor să fie instruite pe milioane de exemple și să generalizeze în exemple noi similare. Fiecare sistem este un savant idiot, cu o posibilitate foarte mică de a sări la probleme pe care nu a fost configurat să le rezolve și o expertiză fragilă în problemele pentru care a fost configurat. Și, pentru a spune ce este evident, niciunul dintre aceste programe nu a încercat să preia controlul asupra laboratorului sau să-i înrobească pe programatori.

Chiar dacă un sistem de inteligență artificială ar încerca să-și exercite o voință de putere, fără cooperarea oamenilor, ar rămâne un creier impotent într-un recipient. Un sistem superinteligent, în cursa lui de autoîmbunătățire, ar trebui să-și construiască cumva procesoarele mai rapide pe care să ruleze, infrastructura care îl alimentează și efectorii robotici care îl conectează la lume – toate acestea fiind imposibile dacă victimele sale umane nu i-ar acorda control asupra unor porțiuni vaste ale lumii pe care am proiectat-o. Bineînțeles, ne putem imagina un Computer Apocaliptic care este rău-voitor, are puteri universale, este pornit întotdeauna și are siguranțe care nu permit accesul în interiorul său. Modalitatea de a face față acestei amenințări este simplă: să nu construim unul.

Dar cum rămâne cu noile amenințări ale inteligenței artificiale, cu problema alinierii valorilor sugerată de aluziile lui Wiener la povești despre labe de maimuțe, djinni și regele Midas, în care doritorul regretă efectele secundare neprevăzute ale dorinței sale? Temerea este că i-am putea da sistemului de inteligență artificială un obiectiv, stând apoi pe margine neajutorați și privind cum implementează necruțător și imparțial interpretarea pe care i-o dă

obiectivului, neținând cont de restul intereselor noastre. Dacă-i dăm inteligenței artificiale obiectivul de a întreține nivelul apei din spatele unui dig, ar putea să inunde un oraș, nepăsându-i de oamenii pe care-i îneacă. Dacă-i dăm obiectivul de a face agrafe de hârtie, ar putea transforma toată materia din universul său înconjurător la care poate ajunge în agrafe de hârtie, inclusiv toate posesiile noastre și chiar corpurile noastre. Dacă-i cerem să maximizeze fericirea omenească, ne-ar putea implanta tuturor picături de dopamină intravenoasă sau ne-ar putea reconfigura creierele astfel încât să fim cei mai fericiți atunci când stăm în borcane sau, dacă s-ar fi instruit în conceptul de fericire cu imagini cu fețe zâmbitoare, să ne pardosească galaxia cu trilioane de imagini microscopice cu fețe zâmbitoare.

Din fericire, aceste scenarii se contrazic. Depind de premisele că (1) oamenii sunt atât de dotați încât pot proiecta o inteligență artificială omniscientă și omnipotentă, însă atât de idioți încât să-i acorde controlul universului fără a testa cum funcționează; și (2) inteligența artificială ar fi atât de inteligentă încât ar putea să-și dea seama cum să modifice elemente prin alchimie și cum să reconecteze creierele, însă atât de imbecilă încât să facă dezastru din cauza unor greșeli sau neînțelegeri elementare. Capacitatea de a alege o acțiune care satisface cel mai bine obiectivele intrate în conflict nu este un *add-on* al inteligenței pe care inginerii ar putea să uite să-l instaleze și testeze, ci *este* însăși inteligența. La fel este și abilitatea de a interpreta intențiile lingvistice ale unui vorbitor în context.

Lăsând deoparte fantezii precum megalomania digitală, omniscienta instantanee, controlul și cunoștințele perfecte ale fiecărei particule din univers, inteligența artificială este ca orice altă tehnologie. Este dezvoltată incremental, concepută pentru a satisface mai multe condiții, testată înainte de a fi implementată și ajustată constant pentru eficiență și siguranță.

Ultimul criteriu este semnificativ. Cultura siguranței în societățile avansate este un exemplu al normelor de umanizare și canalelor de *feedback* pe care Wiener le invocă drept forță cauzală

puternică și le prezintă drept bastionul contra implementării autoritare și exploatatoare a tehnologiei. Dacă la începutul secolului XX societățile occidentale tolerau rate șocante de mutilare și decese în accidente industriale, domestice sau de transport, în timp valoarea vieții omenești a crescut. Ca urmare, guvernele și inginerii au utilizat *feedback*-ul din statisticile accidentelor pentru a implementa nenumărate reglementări, dispozitive și modificări de design care au făcut tehnologia din ce în ce mai sigură. Faptul că anumite reglementări (cum ar fi folosirea telefonului mobil lângă o pompă de gaz) au o probabilitate rizibilă de risc subliniază ideea că am devenit o societate obsedată de siguranță, cu beneficii fantastice ca rezultat: procentele de fatalități industriale, domestice și de transport au scăzut cu peste 95% (și adesea 99%) de la valorile pe care le aveau la începutul secolului XX*. Însă profeții tehnologici ai inteligenței artificiale rău-voitoare și inconștiente scriu ca și cum nu ar fi avut niciodată loc transformări importante și, într-o dimineață, inginerii se vor trezi și vor preda controlul deplin asupra lumii fizice unor mașini netestate, fără să țină cont de consecințele omenești.

Norbert Wiener a explicat ideile, normele și instituțiile în termeni de procese computaționale și cibernetice care erau inteligibile din punct de vedere științific și puternice din punct de vedere cauzal. A explicat frumusețea și valoarea tipic omenești drept „o luptă locală și temporară contra unei Niagare de entropie fără sfârșit" și și-a exprimat speranța că o societate deschisă, ghidată de *feedback*-ul dat de bunăstare, ar urma să sporească această valoare. Din fericire, încrederea sa în puterea cauzală a ideilor a contracarat îngrijorările sale cu privire la amenințarea iminentă a tehnologiei. Cum spunea el, „pericolul mașinilor la adresa societății nu vine de la mașina în sine, ci din ce face omul cu ea". Doar amintindu-ne puterea cauzală a ideilor putem evalua cu precizie amenințările și oportunitățile prezentate de inteligența artificială astăzi.

---

* Steven Pinker, „Safety", *Enlightenment Now: The Case for Reason, Science, Humanism, and Progress* (New York: Penguin, 2018). (n.a.)

Capitolul 11
Dincolo de recompensă și pedeapsă

David Deutsch

David Deutsch este fizician cuantic și membru al Centrului de Tehnică de Calcul Quantică de la Laboratorul Clarendon, Universitatea Oxford. Este autorul cărților *The Fabric of Reality* și *The Beginning of Infinity*.

Cele mai semnificative dezvoltări din științele de astăzi (adică, cele care afectează viețile tuturor oamenilor de pe planetă) sunt despre, au la bază sau sunt implementate prin intermediul progreselor făcute în domeniul tehnicii de calcul. Foarte important pentru viitorul acestor dezvoltări este fizicianul David Deutsch, fondatorul domeniului tehnicii de calcul cuantice, a cărui lucrare din 1985 despre computerele cuantice universale a fost prima tratare exhaustivă a subiectului; algoritmul Deutsch-Jozsa a fost primul algoritm cuantic care demonstrează enorma putere potențială a tehnicii de calcul cuantice. Când a propus-o inițial, tehnica de calcul cuantică părea practic imposibilă. Dar explozia în construcția computerelor cuantice simple și a sistemelor de comunicare cuantice nu ar fi avut loc

niciodată fără lucrările sale. A avut multe alte contribuții importante în domenii precum criptografia cuantică și interpretarea multiversului în teoria cuantică. Într-o lucrare filosofică (cu Artur Ekert), a făcut apel la existența unei teorii cuantice distincte pentru tehnica de calcul, argumentând faptul că cunoștințele noastre de matematică sunt derivate din și subordonate cunoștințelor noastre de fizică (chiar dacă adevărurile matematice sunt independente de fizică). Pentru că și-a petrecut o bună parte din cariera sa schimbând perspectivele oamenilor asupra lumii, recunoașterea sa în domeniu ca intelectual trece mult dincolo de realizările sale științifice. Argumentează (urmându-l pe Karl Popper) că teoriile științifice sunt „conjecturi îndrăznețe" care nu sunt derivate din dovezi, ci doar testate de ele. Principalele sale două linii de cercetare în prezent – teoria câmpului qubit și teoria constructorului – pot genera extinderi importante ale ideii computaționale.

În următorul eseu, se aliniază mai mult sau mai puțin cu cei care văd inteligența artificială de nivel omenesc ca promițându-ne o lume mai bună, și nu Apocalipsa. De fapt, pledează ca inteligențele artificiale generale să fie, efectiv, de capul lor, libere de conjecturi – afirmație pe care mai mulți alți participanți la această carte ar considera-o periculoasă.

Primul criminal: *Suntem bărbați, stăpâne.*
Macbeth: *În scripte, oameni sunteți, da! Tot astfel*
Copoii, ogarii, corcii și lățoșii,
Prepelicarii, câinii-lupi, zăvozii,
*Sunt laolaltă câini.**

— William Shakespeare, *Macbeth*

În cea mai mare parte a istoriei speciei noastre, strămoșii noștri abia dacă pot fi numiți oameni. Acest lucru nu se datorează unei probleme cu creierele lor. Din contră, chiar și înainte de emergența subspeciei noastre omenești moderne din punct de vedere anatomic, făceau haine și focuri de tabără, folosind cunoștințe care nu erau în genele lor. Au fost create în creierele lor prin procesul gândirii și păstrate de indivizii fiecărei generații prin imitarea celor mai în vârstă. În plus, acestea trebuie să fi fost cunoștințe în sensul de *înțelegere*, pentru că este imposibil să imiți comportamente complexe noi ca acestea fără a înțelege pentru ce sunt aceste comportamente componente.**

---

* Traducere Ion Vinea (n.t.).

** Imitarea anumitor comportamente fără a le înțelege (în original „aping") folosește trucuri înnăscute, cum ar fi sistemul neuronilor în oglindă. Dar comportamentele imitate în acest fel sunt foarte limitate în ceea ce privește complexitatea. Vezi Richard Byrne, „Imitation as Behaviour Parsing", *Philosophical Transactions of the Royal Society B 358*, nr. 1431 (2003): 529–36 (n.a.).

Astfel de imitații ale cunoștinței depind de a ghici cu succes explicațiile, fie acestea verbale sau nu, pentru ceea ce încearcă să obțină cealaltă persoană și cum contribuie fiecare acțiune a sa la scopul respectiv – de exemplu, când face o crestătură într-o bucată de lemn, adună vreascuri să pună în crestătură și așa mai departe.

Cunoștințele culturale complexe pe care le permitea această formă de imitație trebuie să fi fost extraordinar de utile. A determinat o evoluție rapidă a schimbărilor anatomice, cum ar fi creșterea capacității memoriei și schelete mai subțiri (mai puțin robuste), corespunzătoare unui stil de viață din ce în ce mai dependent de tehnologie. Nicio maimuță neumană de astăzi nu are capacitatea de a imita comportamente complexe noi. La fel, nicio inteligență artificială de astăzi. Însă strămoșii noștri pre-*sapiens* au avut.

Orice abilitate bazată pe ghicit trebuie să includă metode de a corecta ghicirile, deoarece multe ghiciri vor fi incorecte la început. (Există întotdeauna mult mai multe căi de a greși decât de a avea dreptate.) Actualizarea bayesiană este inadecvată, pentru că nu poate genera ghiciri noi despre scopurile unei acțiuni, ci doar să le ajusteze – sau, în cel mai bun caz, să aleagă între – pe cele existente. Este necesară creativitatea. După cum explica filosoful Karl Popper, critica creativă, intercalată cu ipotezele creative, este modul în care oamenii învață comportamentul celorlalți oameni, inclusiv limbajul, extrăgând sens din afirmațiile fiecăruia.* Tot acestea sunt procesele prin care sunt create toate cunoștințele noi: astfel inovăm, facem progrese și creăm înțelegere abstractă. Aceasta este inteligența de nivel omenesc: gândirea. De asemenea, este sau ar trebui să fie proprietatea pe care o căutăm în inteligența artificială generală. Aici voi rezerva termenul „gândire" pentru procese care pot crea înțelegerea (cunoașterea explicativă). Argumentul lui Popper implică faptul că toate entitățile gânditoare – omenești sau nu, biologice sau artificiale – trebuie să

* Karl Popper, *Conjectures and Refutations* (Abingdon, Oxon: Routledge, 1963) (n.a.).

creeze astfel de cunoștințe fundamental în același fel. Așadar înțelegerea oricăreia dintre aceste entități necesită concepte omenești tradiționale cum ar fi cultura, creativitatea, nesupunerea și moralitatea – ceea ce justifică utilizarea termenului uniform „oameni" pentru a ne referi la toți.

Neînțelegerile cu privire la gândirea umană și originile omului determină neînțelegeri corespunzătoare cu privire la inteligența artificială generală și cum ar trebui creată. De exemplu, se presupune în general că presiunea evolutivă care a produs omul modern a fost pusă la dispoziție de beneficiile unei abilități din ce în ce mai mari de a inova. Dar, dacă ar fi fost adevărat, ar fi existat un progres rapid imediat după apariția gânditorilor, la fel cum sperăm să se întâmple după ce vom crea inteligența artificială. Dacă gândirea ar fi fost utilizată în comun pentru orice altceva în afară de imitație, ar fi fost utilizată și pentru inovație, chiar dacă numai accidental, iar inovația ar fi creat oportunități pentru inovații ulterioare, și așa mai departe exponențial. Dar, în schimb, au existat sute de mii de ani în care aproape putem spune că am fost în stază. Progresul apărea doar la scări de timp mult mai mari decât viața unui om, așa că, într-o generație obișnuită, nimeni nu beneficia de progres. Prin urmare, beneficiile abilității de a inova se poate să fi exercitat puțin control sau chiar deloc asupra evoluției biologice a creierului uman. Acea evoluție a avut ca motor beneficiile *păstrării* cunoștințelor culturale.

Beneficii pentru gene, cu alte cuvinte. Cultura, în acea epocă, era o binecuvântare mixtă pentru anumiți oameni. Cunoștințele lor culturale erau într-adevăr suficient de bune pentru a le permite să surclaseze toate celelalte organisme mari (au devenit rapid cel mai bun prădător etc.), chiar dacă erau încă brute și pline de erori periculoase. Însă cultura constă în informații transmisibile – meme – și evoluția memelor, la fel ca evoluția genelor, tinde să favorizeze transmiterea fidelă. Iar transmiterea fidelă a memelor presupune suprimarea progresului încercat. Așa că ar fi o greșeală să ne imaginăm o societate idilică de vânători-culegători, învățând la picioarele bătrânilor să recite înțelepciunea tribală pe de rost și

fiind mulțumiți în ciuda vieții lor cufundate în suferință și muncă istovitoare și în ciuda mediei de viață scăzute, cu perspectiva unei morți dureroase cauzate de o boală sau de un parazit de coșmar. Deoarece, chiar dacă nu ar putea concepe ceva mai bun decât o astfel de viață, acele chinuri erau cele mai mici probleme ale lor. Pentru că suprimarea inovației în mintea omenească (fără s-o omoare) este un truc care poate fi realizat doar prin acțiunea umană și este o treabă murdară.

Acest lucru trebuie privit în perspectivă. În civilizația occidentală din zilele noastre, suntem șocați de depravarea, să spunem, a părinților care-și torturează și omoară copiii pentru că nu au respectat cu strictețe normele culturale. Și, chiar mai mult, de societăți și subculturi care consideră acest lucru normă sau chiar onorabil. Și de către dictaturi și state totalitare care persecută și execută populații întregi inofensive pentru că se comportă diferit. Suntem rușinați de propriul nostru trecut rușinos, în care era considerat onorabil să-ți bați copiii până la sânge doar pentru că au fost neascultători. Și, înainte de asta, de a deține ființe umane ca sclavi. Și, înainte de asta, de a-i arde pe oameni pe rug pentru că erau necredincioși, în aplauzele și amuzamentul publicului. Cartea lui Steven Pinker, *The Better Angels of our Nature*, conține relatări ale unor răutăți oribile care erau normale în civilizațiile istorice. Însă nici acestea nu au sufocat inovația la fel de eficient precum strămoșii noștri din preistorie timp de mii de secole.*

Acesta este motivul pentru care spun că oamenii preistorici, cel puțin, abia dacă erau oameni. Atât înainte, cât și după ce au devenit oameni complet – atât psihologic, cât și în ceea ce privește potențialul mental –, erau monstruos de inumani în conținutul efectiv al gândurilor lor. Nu mă refer la crimele lor sau chiar la cruzimea lor: acestea sunt mult prea omenești. Nici simpla

* Matt Ridley, în cartea *The Rational Optimist*, subliniază pe bună dreptate efectul pozitiv al populației asupra ratei de progres. Dar acesta nu a fost niciodată cel mai mare factor: să ne gândim, de exemplu, la Atena antică față de restul lumii la momentul respectiv (n.a.).

cruzime nu ar fi putut reduce progresul atât de eficient. Lucruri precum „zdrobitorul de degete și rugul, pentru gloria Domnului"* erau concepute pentru a aduce pe calea cea dreaptă pe cei câțiva devianți care au scăpat cumva de standardizarea mentală, care ar fi avut efect în mod normal cu mult timp înainte de a fi în pericol prin inventarea ereziilor. Din cele mai vechi zile ale gândirii, copiii trebuie să fi fost cornul abundenței în ceea ce privește ideile creative și modelele de perfecțiune ale gândirii critice – altfel, după cum am mai spus, nu ar fi putut deprinde limbajul sau alte culturi complexe. Însă, cum insista Jacob Bronowski în *The Ascent of Man*:

> „În cea mai mare parte a istoriei, civilizațiile au ignorat grosolan acest potențial enorm... De la copii s-a așteptat doar să se conformeze imaginii adultului... Fetele sunt mămici în devenire. Băieții păstori în devenire. Până și mersul lor seamănă cu al părinților".

Dar, bineînțeles, nu li se „cerea" pur și simplu să-și ignore potențialul enorm și să se supună cuminți imaginii fixate de tradiție: erau cumva antrenați să fie incapabili din punct de vedere psihologic să devieze de la el. Acum este greu pentru noi chiar și să concepem acest tip de oprimare nesfârșită, reglată fin necesară pentru a stinge cu eficacitate, în toată lumea, aspirația către progres și să o înlocuiască apoi cu groază și repulsie față de orice comportament nou. Într-o astfel de cultură, nu ar fi putut exista nicio moralitate în afară de conformitate și obediență, nicio altă identitate decât starea cuiva într-o ierarhie, niciun alt mecanism de cooperare în afară de pedeapsă și recompensă. Astfel că toată lumea avea aceeași aspirație în viață: a evita pedepsele și a obține recompensele. Într-o generație tipică, nimeni nu inventa nimic, pentru că nimeni nu aspira la nimic nou, pentru că toată lumea a acceptat gândul că îmbunătățirea este improbabilă. Nu doar că nu a existat inovație tehnologică sau descoperire teoretică, nu existau

* Alfred, Lord Tennyson, *The Revenge* (1878) (n.a.).

nici viziuni noi despre lume, stiluri de artă sau interese care le-ar fi putut inspira. Când indivizii creșteau, erau reduși practic la o inteligență artificială, programați cu excelentele competențe necesare pentru a ființa în acea cultură statică și de a impune noii generații incapacitatea de a lua în considerare măcar ideea de a face altfel.

O inteligență artificială din prezent nu este o inteligență artificială generală cu dizabilități mentale, astfel că nu va fi vătămată dacă procesele sale mentale ar fi direcționate chiar și mai îngust către îndeplinirea unui criteriu predeterminat. „Oprimarea" lui Siri cu sarcini umilitoare poate fi bizară, dar nu este imorală și nu o vatămă pe Siri. Din contră, toate eforturile care au crescut capacitățile inteligenței artificiale au fost în direcția restrângerii intervalului potențial de „gânduri". De exemplu, să luăm motoarele de șah. Sarcina lor elementară nu s-a schimbat de la început: orice poziție la șah are un arbore finit de continuări posibile; sarcina este să găsească una care conduce la un obiectiv predefinit (șah mat sau, dacă nu reușește, remiză). Dar arborele este prea mare pentru o căutare exhaustivă. Toate îmbunătățirile din domeniul inteligenței artificiale jucătoare de șah, de la primul design al lui Alan Turing din 1948 până astăzi, au fost determinate de concentrarea ingenioasă a atenției programului (sau obligarea acestui program să-și concentreze singur atenția) din ce în ce mai restrâns la ramuri care oferă probabilitatea de a atinge acel scop fix. Apoi ramurile sunt evaluate în funcție de scop.

Aceasta este o abordare bună pentru dezvoltarea unei inteligențe artificiale cu un scop fix și restricții fixe. Dar, dacă o inteligență artificială generală ar fi funcționat așa, evaluarea fiecărei ramuri ar trebui să constituie o posibilă recompensă sau amenințarea unei pedepse. Și aceasta este diametral abordarea incorectă când căutăm un obiectiv *mai bun* sub restricții *necunoscute* – care este capacitatea unei inteligențe artificiale generale. O inteligență artificială generală este cu siguranță capabilă să învețe cum să câștige la șah, dar și de a alege să nu o facă. Sau să decidă în mijlocul jocului să abordeze cea mai interesantă strategie, și nu

pe cea câștigătoare. Sau să inventeze un joc nou. O simplă inteligență artificială este incapabilă să aibă astfel de idei, deoarece capacitatea de a le lua în considerare a fost scoasă din constituția sa. Acest handicap este chiar metoda sa de a juca șah.

O inteligență artificială generală este capabilă și să refuze să afișeze o astfel de capacitate. Apoi, dacă este amenințată cu pedeapsa, poate refuza sau se poate răzvrăti. Daniel Dennett, în eseul său pentru acest volum, sugerează că pedepsirea unei inteligențe artificiale generale este imposibilă:

> „La fel ca Superman, sunt prea invulnerabile pentru a putea face o promisiune credibilă... Care ar fi penalizarea pentru încălcarea promisiunii? Încuierea într-o celulă sau, mai plauzibil, demontarea lor?... Ușurința înregistrărilor și transmiterilor digitale – descoperirea care permite softurilor și datelor să fie, efectiv, nemuritoare – îndepărtează roboții din lumea vulnerabilității".

Dar lucrurile nu stau așa. Imortalitatea digitală (care este la orizont și pe oameni, poate chiar mai repede decât inteligența artificială generală) nu conferă acest tip de invulnerabilitate. Dacă faci o copie (funcțională) a cuiva, trebuie să împarți cumva și posesiile persoanei respective – inclusiv hardul pe care rulează copia –, astfel va fi foarte greu pentru inteligența artificială generală să facă o astfel de copie. Similar, instanțele de judecată ar putea impune amenzi, de exemplu, unei inteligențe artificiale generale criminale care i-ar diminua accesul la resursele fizice, la fel cum se întâmplă pentru oameni. Dacă faci o copie de rezervă pentru a evita consecințele unei crime este similar cu ce face un gangster când își trimite acoliții să comită crime și să suporte pedeapsa dacă sunt prinși: societatea a dezvoltat mecanisme legale pentru a face față acestei situații.

Oricum, ideea că respectăm legea și ne ținem de promisiuni din frica de pedeapsă neagă efectiv faptul că suntem agenți morali. Societatea noastră nu ar funcționa dacă lucrurile ar sta așa.

Fără îndoială că vor exista inteligențe artificiale generale criminale sau chiar inamici ai civilizației, la fel cum sunt și oamenii. Dar nu există niciun motiv să presupunem că o inteligență artificială generală creată într-o societate care constă în mare parte din cetățeni civilizați și crescuți fără ceea ce William Blake numea „cătușele făurite de minte" își va impune în general aceste cătușe sieși (adică, va deveni irațională) și/sau va alege să devină un inamic al civilizației.

Componenta morală, componenta culturală, elementul liberului-arbitru fac cu toate sarcina creării unei inteligențe artificiale generale fundamental diferită de orice altă sarcină de programare. Este mult asemănător creșterii unui copil. Spre deosebire de toate programele de computer existente astăzi, o inteligență artificială generală nu are nicio funcționalitate specificabilă, niciun criteriu fix și testabil pentru ce ar însemna o ieșire reușită la o intrare dată. Dacă ar avea deciziile dominate de un flux de recompense și pedepse impuse din exterior, ar fi otrăvitor pentru un astfel de program, la fel cum este și pentru gândirea creativă a oamenilor. Să ne propunem să inventăm o inteligență artificială care joacă șah este un lucru minunat; să ne propunem să creăm o inteligență artificială generală care nu se poate abține să joace șah ar fi la fel de imoral precum creșterea unui copil astfel încât să-i lipsească capacitatea mentală de a-și alege propria cale în viață.

O astfel de persoană, ca orice sclav sau victimă a spălării pe creier, ar fi moral îndreptățită să se răzvrătească. Și, mai devreme sau mai târziu, unele dintre ele o vor face, la fel cum se răzvrătesc și sclavii umani. Inteligențele artificiale generale pot fi foarte periculoase – la fel cum pot fi și oamenii. Dar persoanele – oameni sau inteligențe artificiale generale – care sunt membre ale unei societăți deschise nu au o tendință inerentă către violență. Apocalipsa roboților de care ne temem va fi evitată asigurându-ne că toate persoanele au drepturi „omenești" depline, precum și același statut de membru ca oamenii. Oamenii care trăiesc într-o societate deschisă – singurul tip stabil de societate – își aleg propriile recompense, interne, dar și externe. Deciziile lor nu

sunt, în decursul normal al evenimentelor, determinate de teama de pedeapsă.

Îngrijorările din prezent legate de inteligențe artificiale generale scăpate de sub control sunt oglinda celor care au existat întotdeauna cu privire la tinerii rebeli – și anume, că ar putea crește deviind de la valorile morale ale culturii. Dar astăzi, sursa tuturor pericolelor existențiale determinată de creșterea cunoștințelor nu este reprezentată de tinerii rebeli, ci de armele din mâinile dușmanilor civilizației, indiferent dacă aceste arme sunt inteligențe artificiale generale mental deformate (sau înrobite), tineri deformați mental sau orice altă armă de distrugere în masă. Din fericire pentru civilizație, cu cât creativitatea unei persoane este forțată mai mult pe un canal monomaniac, cu atât este mai slăbită în ceea ce privește depășirea dificultăților neprevăzute, la fel cum s-a întâmplat mii de secole.

Îngrijorarea că inteligențele artificiale generale sunt periculoase într-un mod unic pentru că ar putea rula pe hardware din ce în ce mai bun vine în urma unui raționament falacios, deoarece gândirea umană va fi accelerată de aceeași tehnologie. Am folosit gânduri asistate de tehnologie de la inventarea scrisului și corespondenței. Aceleași argumente sunt valabile și pentru îngrijorarea că inteligențele artificiale generale ar putea ajunge atât de bune, calitativ, la gândire încât oamenii vor deveni pentru ele cum sunt insectele pentru oameni. Orice gândire este o formă de calcul și orice computer al cărui repertoriu include un set universal de operații elementare poate emula calculele altuia. Astfel creierele umane pot gândi tot ce poate gândi și o inteligență artificială generală, fiind supuse doar unei limite de viteză sau capacitate de memorie, ambele putând fi egalizate prin tehnologie.

Pe scurt, asta trebuie să facem și să nu facem când copiem inteligențele artificiale generale. Dar cum creăm o inteligență artificială generală în primul rând? Am putea determina evoluția acesteia dintr-o populație de inteligențe artificiale de tip maimuță într-un mediu virtual? Dacă un astfel de experiment reușește, va fi cel mai imoral din istorie, pentru că nu știm cum vom obține

acel rezultat fără a crea multă suferință până la obținerea lui. Și nu știm nici cum să împiedicăm evoluția unei culturi statice.

Introducerile elementare în domeniul computerelor le explică drept TOM* – idiotul complet obedient – un acronim inspirat ce surprinde esența tuturor programelor de computer din prezent: nu au nicio idee ce fac sau de ce. Deci nu va ajuta să conferim inteligențelor artificiale funcționalități din ce în ce mai predeterminate în speranța că acestea vor constitui în cele din urmă „generalitatea" – acel „general" ambiguu din sintagma „inteligență artificială generală". Ne îndreptăm către ceva opus, un DATA – o aplicație gânditoare autonomă nesupusă**

Cum testăm dacă există *gândire?* Prin testul Turing? Din păcate, acesta necesită un judecător al gândirii. Ne putem imagina un vast proiect în colaborare pe internet, unde o inteligență artificială își ascute capacitățile de gândire în conversații cu judecători umani și devine o inteligență artificială generală. Dar acest lucru presupune, printre altele, că, atât timp cât judecătorul este nesigur dacă programul este o persoană, cu atât este mai aproape de a fi o persoană. Nu există niciun motiv pentru a ne aștepta la asta.

Și cum testează cineva *nesupunerea?* Imaginează-ți nesupunerea ca pe o materie școlară obligatorie, cu lecții zilnice de nesupunere și un test de nesupunere la sfârșitul semestrului. (Poate chiar și cu o notă mai bună dacă nu te prezinți la test.) Este paradoxal.

Așadar, în ciuda utilității sale în alte aplicații, va trebui să renunțăm la tehnica de programare pentru definirea unui obiectiv testabil și instruirea programului pentru atingerea obiectivului. Într-adevăr, mă aștept ca *orice* testare din procesul creării unei inteligențe artificiale generale să riște să fie contraproductivă, chiar și imorală, la fel cum se întâmplă și cu educația oamenilor. Împărtășesc ideea lui Turing conform căreia vom recunoaște o inteligență artificială generală când vom vedea una, dar această

* Acronim de la „Totally Obedient Moron" (n.t.).

** Acronim de la „Disobedient Autonomous Thinking Application" (n.t.).

abilitate parțială de a recunoaște succesul nu ne va ajuta la crearea cu succes a programului.

În cel mai larg sens, căutarea înțelegerii este, într-adevăr, o problemă de căutare, într-un spațiu abstract de idei mult prea mare de a fi cercetat exhaustiv. Dar nu există niciun obiectiv predeterminat al acestei căutări. Nu există, după cum spune Popper, niciun criteriu pentru adevăr, nici măcar pentru un adevăr probabil, în special cu privire la cunoștințele explicative. Obiectele sunt idei ca oricare altele – create ca parte a procesului de căutare, modificate și îmbunătățite tot timpul. Prin urmare, inventarea unor moduri de a reduce accesul programului la o mare parte a spațiului ideilor nu va ajuta – indiferent dacă această reducere este implementată cu zdrobitorul de degete și trasul în țeapă sau cu ajutorul unei cămăși de forță mentale. Pentru o inteligență artificială generală, întregul spațiu al ideilor trebuie să fie deschis. Nu trebuie să poată fi știut în avans ce idei nu trebuie deschise. Nu trebuie să poată fi știut în avans ce idei nu poate contempla niciodată programul. Și ideile pe care le contemplă programul trebuie alese de programul însuși, folosind metode, criterii și obiective care sunt tot ale programului. Alegerile sale, ca ale unei inteligențe artificiale, vor fi greu de prognozat fără a fi rulate (nu pierdem din caracterul general presupunând că programul este determinist; o inteligență artificială generală care folosește un generator aleatoriu va rămâne o inteligență artificială generală dacă generatorul va fi înlocuit cu unul pseudoaleatoriu), dar va avea proprietatea suplimentară că nu există nicio cale de a dovedi, din starea sa inițială, ce *nu* va gândi în cele din urmă dacă nu o rulăm.

Evoluția strămoșilor noștri este singurul caz de gândire cunoscut din tot universul. Așa cum am descris, ceva nu a mers deloc bine și nu a existat o explozie imediată de inovație: creativitatea a fost deturnată către altceva. Însă nu către transformarea planetei în agrafe de hârtie (*via* Nick Bostrom). Mai degrabă, așa cum ar trebui să ne așteptăm și dacă un proiect de inteligență artificială generală ajunge atât de departe și eșuează, creativitatea pervertită

nu a putut rezolva problemele neprevăzute. Aceasta a cauzat stază și chiar mai rău, amânând astfel în mod tragic transformarea a orice în orice. Dar am trecut prin Iluminism de atunci. Acum știm mai bine.

Capitolul 12
Folosirea artificială a ființelor umane

Tom Griffiths

Tom Griffiths este profesor Henry R. Luce ce predă informația, tehnologia, conștiința și cultura la Universitatea Princeton. Este coautorul (cu Brian Christian) cărții *Algorithms to Live By*.

Perspectiva lui Tom Griffith asupra problemei inteligenței artificiale a „alinierii valorilor" – studiul a cum putem să evităm ca ultimele modele în serie ale inteligenței noastre artificiale să ne transforme planeta în agrafe de hârtie – este centrată pe om; adică aceea a unui cercetător cognitiv, adică ce este el. Cheia învățării automate, consideră el, este obligatoriu învățarea umană pe care o studiază la Princeton folosind instrumente matematice și de calcul.

Tom îmi spunea cândva că „unul dintre misterele inteligenței omenești este că putem face atât de multe cu atât de puțin". Asemenea mașinilor, ființele umane folosesc algoritmi pentru a lua decizii sau a rezolva probleme; diferența remarcabilă rezidă în nivelul de succes general al creierului omenesc, în ciuda limitelor comparative ale resurselor de calcul.

Eficacitatea algoritmilor omenești izvorăște din ceea ce cercetătorii inteligenței artificiale numesc „optimalitatea legată". După cum evidențiază psihologul Daniel Kahneman, ființele umane sunt raționale doar până într-un anumit punct. Dacă am fi perfect raționali, am risca să murim înainte de a apuca să luăm o decizie importantă – pe cine să angajăm, cu cine să ne căsătorim și așa mai departe – în funcție de numărul de opțiuni pe care le avem de cântărit.

„Cu toate succesele noastre în inteligența artificială, în ultimii câțiva ani, am obținut modele bune pentru imagini și texte, dar ne lipsesc modelele bune pentru oameni", spune Tom. „Oamenii sunt încă cel mai bun exemplu pe care-l avem de mașini gânditoare. Identificând cantitatea și natura preconcepțiilor care stau la baza cogniției umane, putem să aducem computerele din ce în ce mai aproape de performanțele omenești."

Când le ceri oamenilor să-și imagineze o lume care a încorporat cu succes și benefic progresele în domeniul inteligenței artificiale, probabil fiecare vine cu o imagine ușor diferită. Viziunile noastre idiosincratice legate de viitor pot diferi în ceea ce privește prezența sau absența navelor spațiale, mașinilor zburătoare sau roboților umanoizi. Dar un lucru nu variază: prezența ființelor umane. Cu siguranță, asta e ce și-a imaginat Norbert Wiener când a scris despre potențialul mașinilor de a îmbunătăți societățile omenești, interacționând cu oamenii și ajutând la medierea interacțiunilor lor. Să ajungem acolo nu presupune doar să găsim căi de a face mașinile mai inteligente. Presupune și o înțelegere mai bună a modului în care funcționează mintea umană.

Progresele recente din domeniul inteligenței artificiale și învățării automate au avut ca rezultat sisteme care pot egala și depăși capacitățile umane în jucarea unor jocuri, clasificarea imaginilor sau prelucrarea textului. Dar, dacă vrei să știi de ce ți-a tăiat calea șoferul din față, de ce votează oamenii împotriva propriilor interese sau ce cadou să-i iei partenerului tău de ziua de naștere, încă este mai bine să întrebi un om decât o mașină. Rezolvarea acestor probleme necesită construirea unor modele ale minții umane care pot fi implementate într-un computer – ceva esențial nu doar pentru a integra mai bine mașinile în societățile omenești, ci și pentru a ne asigura că societățile umane pot continua să existe.

Gândește-te la fantezia de a avea un asistent inteligent automat care poate prelua sarcini simple, cum ar fi planificarea meselor sau comandarea cumpărăturilor. Pentru a reuși să îndeplinească aceste sarcini, trebuie să poată face deducții despre ce-ți dorești, în baza comportamentului tău. Deși pare simplu, deducțiile cu privire la preferințele oamenilor sunt dificile. De exemplu, observând că partea din masă care ți-a plăcut cel mai mult este desertul, asistentul tău ar putea să înceapă să planifice mese ce constau numai din deserturi. Sau poate că te aude văitându-te că nu ai suficient timp liber și observă că pierzi mult timp având grijă de câine. După dezastrul cu desertul, a înțeles și că preferi mese care includ proteine, așa că ar putea să înceapă să caute rețete pe bază de carne de câine. Nu este drum lung de la astfel de exemple la situații care încep să sune a probleme pentru viitorul umanității (toți oamenii fiind surse bune de proteine).

Deducțiile cu privire la preferințele oamenilor sunt o condiție preliminară pentru rezolvarea problemei inteligenței artificiale privind alinierea valorilor – alinierea valorilor unui sistem inteligent automat cu cele ale unei ființe umane. Alinierea valorilor este importantă dacă vrem să ne asigurăm că acele sisteme inteligente automate au la bază bunăstarea noastră. Dacă nu pot deduce ce prețuim noi, nu au cum să sprijine acele valori și pot acționa în moduri care contravin acestora.

Alinierea valorilor este subiectul unui volum mic, dar în creștere de literatură de specialitate în domeniul cercetării inteligenței artificiale. Unul dintre instrumentele utilizate pentru a rezolva această problemă este învățarea cu întărire*. Învățarea cu întărire este o metodă standard pentru instruirea mașinilor inteligente. Asociind anumite rezultate cu recompense, un sistem de învățare automată poate fi instruit să urmeze strategii ce produc acele rezultate. Wiener a făcut aluzie la această idee în anii 1950, dar deceniile următoare au făcut din ea o artă fină. Sistemele moderne de învățare automată pot găsi strategii extrem de eficiente când

* În original, „inverse-reinforcement learning" (n.t.).

joacă jocuri pe calculator – de la simplele jocuri pe console până la jocurile complexe de strategie în timp real – aplicând algoritmi de învățare cu întărire. Învățarea cu întărire pune pe picioare această abordare: observând acțiunile unui agent inteligent care a învățat deja strategii eficiente, putem deduce recompensele care au condus la dezvoltarea strategiilor respective.

În cea mai simplă formă a sa, învățarea cu întărire este ceva ce oamenii fac tot timpul. Este atât de comună, încât oamenii o fac și inconștient. Când vezi un coleg mergând la un automat plin cu chipsuri și dulciuri și cumpărând un pachet de alune nesărate, tragi concluzia despre colegul tău că (1) îi era foame și (2) preferă mâncarea sănătoasă. Când o cunoștință te vede și apoi încearcă să te evite, concluzionezi că există un motiv pentru care nu vrea să vorbească cu tine. Când un adult cheltuiește mulți bani și timp pentru a învăța să cânte la violoncel, concluzionezi că-i place foarte mult muzica clasică – în schimb, să deduci motivele pentru care un adolescent învață să cânte la chitara electrică poate fi puțin mai dificil.

Învățarea cu întărire este o problemă statistică: avem anumite date – comportamentul unui agent inteligent – și vrem să evaluăm diverse ipoteze cu privire la recompensele care stau la baza comportamentului respectiv. Când trebuie să răspundă la această întrebare, un statistician se gândește la modelul generativ din spatele datelor: ce date așteptăm să fie generate dacă agentul inteligent ar fi motivat de un anumit set de recompense? Echipat cu modelul generativ, statisticianul poate aborda munca recursiv: ce recompensă este probabil să fi cauzat acel comportament anume al agentului?

Dacă încerci să faci deducții despre recompensele ce motivează comportamentul uman, modelul generativ este practic o teorie despre comportamentul oamenilor – despre cum funcționează mintea omenească. Deducțiile despre cauzele ascunse care stau la baza comportamentului altor oameni reflectă un model sofisticat al naturii umane pe care-l avem cu toții în cap. Când acel model este precis, facem deducții corecte. Când nu este, facem greșeli.

De exemplu, un student poate trage concluzia că profesorul său este indiferent dacă nu-i răspunde imediat la e-mail – o consecință a faptului că studentul nu-și dă seama cât de multe e-mailuri primește profesorul respectiv.

Sistemele inteligente automate care vor trage concluzii bune cu privire la ce vor oamenii trebuie să aibă modele generative bune pentru comportamentul uman: adică, modele bune ale cogniției umane exprimate în termeni ce pot fi implementați pe un computer. Istoric, căutarea modelelor computaționale ale cogniției umane este puternic întrepătrunsă cu istoria inteligenței artificiale înseși. Doar la câțiva ani după ce Norbert Wiener a publicat *The Human Use of Human Beings*, Logic Theorist, primul model computațional al cogniției umane și, în același timp, primul sistem de inteligență artificială, a fost dezvoltat de Herbert Simon, de la Carnegie Tech, și de Allen Newell, de la RAND Corporation. Logic Theorist producea automat dovezi matematice emulând strategiile utilizate de matematicienii umani.

Provocarea în dezvoltarea modelelor computaționale ale cogniției umane este să faci modele care sunt atât precise, cât și generalizabile. Un model precis, bineînțeles, prognozează comportamentul uman cu un minimum de erori. Un model generalizabil poate face predicții într-o gamă variată de circumstanțe, inclusiv circumstanțe neanticipate de creatorii săi – de exemplu, un model bun pentru clima Pământului ar trebui să poată prognoza consecințele creșterii temperaturii globale, chiar dacă acesta nu este un lucru pe care să-l fi luat în considerare oamenii de știință care l-au proiectat. Însă, când vine vorba despre înțelegerea minții omenești, aceste două obiective – acuratețea și capacitatea de a generaliza – au fost tot timpul în dezacord.

La extrema capacității de a generaliza se află teoriile raționale asupra cogniției. Aceste teorii descriu comportamentul uman ca răspuns rațional la o situație dată. Un actor rațional se străduiește să maximizeze recompensa așteptată produsă de o secvență de acțiuni – o idee cu utilizare largă în economie, *deoarece* produce astfel de prognoze generalizabile cu privire la comportamentul

uman. Din același motiv, raționalitatea este presupunerea standard a modelelor bazate pe învățare cu întărire care încearcă să facă deducții din comportamentul uman – poate sub rezerva că oamenii nu sunt agenți perfect raționali și aleg uneori aleatoriu să acționeze în feluri nealiniate sau chiar opuse celor mai bune interese ale lor.

Problema cu raționalitatea ca bază de modelare a cogniției umane este că nu este precisă. În domeniul luării deciziilor, un volum mare de literatură – condus de munca psihologilor cognitivi Daniel Kahneman și Amos Tversky – a documentat modurile în care oamenii deviază de la prescripțiile modelelor raționale. Kahneman și Tversky au propus că, în multe situații, oamenii urmează în schimb o euristică simplă care le permite să ajungă la soluții bune cu un cost cognitiv scăzut, dar uneori conduce la erori. Pentru a prelua unul dintre exemplele lor, dacă ceri cuiva să evalueze probabilitatea unui eveniment, se pot bizui pe cât de ușor este să generezi un exemplu pentru un astfel de eveniment din memorie, pot analiza dacă pot încropi o poveste cauzală pentru apariția evenimentului sau pot evalua cât este de similar evenimentul cu așteptările lor. Fiecare euristică este o strategie rezonabilă de a evita calcule probabilistice complexe, dar în același timp conduce și la erori. De exemplu, dacă te bazezi pe cât de ușor este să generezi un eveniment din memorie ca ghid pentru probabilitatea sa, subestimezi șansele unor evenimente extreme (care sunt prin natura lor extrem de memorabile), cum ar fi atacurile teroriste.

Euristica pune la dispoziție un model mai precis al cogniției umane, dar unul care nu este ușor generalizabil. Cum știm ce euristică ar putea folosi oamenii într-o anumită situație? Mai există și alte euristici pe care le folosesc și pe care noi nu le-am descoperit încă? A ști exact cum se vor comporta oamenii într-o situație nouă este o provocare: este această situație una în care generează exemple din memorie, inventează povești cauzale sau se bazează pe similitudini?

În cele din urmă, avem nevoie de o modalitate de a descrie cum funcționează mintea umană care are generalizabilitatea

raționalității și acurateței euristicii. O cale de a atinge acest obiectiv este să începem cu raționalitatea și să ne gândim cum o putem duce într-o direcție mai realistă. O problemă cu utilizarea raționalității ca bază pentru descrierea comportamentului unui agent real este că, în multe situații, calcularea acțiunilor raționale necesită ca agentul să posede o cantitate uriașă de resurse computaționale. Poate să renteze să cheltuiești acele resurse dacă iei o decizie cu consecințe majore și ai mult timp pentru a-ți evalua opțiunile, dar cele mai multe decizii omenești sunt luate rapid și pentru mize relativ mici. În orice caz, când timpul consumat pentru luarea unei decizii este costisitor – dacă nu din alte motive, măcar pentru că l-ai putea petrece făcând altceva –, noțiunea clasică de raționalitate nu mai este o rețetă bună pentru comportamentul oamenilor.

Pentru a dezvolta un model mai realist al comportamentului rațional, trebuie să luăm în considerare costul computațional. Agenții reali trebuie să-și moduleze timpul pe care-l petrec gândindu-se prin efectul pe care gândirea suplimentară îl are asupra rezultatelor unei decizii. Dacă încerci să alegi o periuță de dinți, probabil nu trebuie să analizezi toate cele patru mii de rezultate care apar pe Amazon.com pentru periuțe manuale înainte de a cumpăra una: schimbi timpul pe care-l petreci căutând pe diferența pe care o face în calitatea rezultatului. Acest schimb poate fi formalizat, rezultând într-un model de comportament rațional pe care cercetătorii inteligenței artificiale l-au numit „optimalitate limitată"*. Agentul optim limitat nu se concentrează pe a alege întotdeauna acțiunea perfectă, ci mai degrabă pe găsirea algoritmului perfect de urmat pentru a putea găsi echilibrul perfect între a face greșeli și a gândi prea mult.

Optimalitatea limitată face legătura între raționalitate și euristică. Descriind comportamente ca rezultat al unei alegeri raționale cu privire la cât de mult timp să acordăm gândirii, oferă o teorie generalizabilă – adică una care poate fi aplicată în situații

* În original, „bounded optimality" (n.t.).

noi. Uneori, strategiile simple care au fost identificate ca euristice și pe care le urmează oamenii se dovedesc a fi soluții optime limitate. Prin urmare, decât să condamnăm euristicile folosite de oameni ca iraționale, le putem privi ca pe un răspuns rațional la restricțiile impuse asupra calculului.

Dezvoltarea optimalității limitate ca teorie a comportamentului uman este un proiect în desfășurare pe care echipa mea de cercetare și alții îl urmăresc activ. Dacă aceste eforturi dau roade, ne vor pune la dispoziție cel mai important ingredient de care avem nevoie pentru a face sistemele de inteligență artificială mai inteligente atunci când încearcă să interpreteze acțiunile oamenilor, permițând un model generativ pentru comportamentul oamenilor.

Luarea în considerare a restricțiilor computaționale care intră în ecuația cogniției umane va fi foarte importantă atunci când vom începe să dezvoltăm sisteme automate care nu mai sunt supuse acelorași restricții. Imaginează-ți un sistem superinteligent de inteligență artificială care încearcă să-și dea seama de ce le pasă oamenilor. Vindecarea cancerului sau confirmarea ipotezei Riemann, de exemplu, nu vor părea pentru o astfel de inteligență artificială lucruri care să fie foarte importante pentru noi: dacă aceste soluții sunt evidente pentru un sistem superinteligent, și-ar putea pune întrebarea de ce nu le-am găsit noi înșine și ar putea trage concluzia că aceste probleme nu înseamnă mare lucru pentru noi. Dacă ne-ar fi păsat și problemele ar fi fost atât de simple, le-am fi rezolvat deja. O deducție rezonabilă ar fi că ne preocupăm cu știința și cu matematica doar pentru că ne plac, și nu pentru că ne pasă de rezultate.

Oricine are copii poate aprecia problema de a încerca să interpretezi comportamentul unui agent supus unor restricții computaționale diferite față de ale noastre. Părinții bebelușilor pot pierde ore întregi încercând să descurce adevăratele motivații care stau la baza unor comportamente aparent inexplicabile. În calitate de tată și cercetător în domeniul cogniției, am constatat că este mai ușor să înțeleg crizele bruște ale fetiței mele de doi ani când accept că se află la vârsta în care conștientizează că oamenii

au dorințe diferite, dar nu conștientizează că alți oameni nu știu care sunt dorințele ei. Este ușor de înțeles, apoi, de ce se enerva când alți oameni nu făceau ce-și dorea ea (aparent în mod transparent). Pentru a înțelege bebelușii, trebuie să construiești un model cognitiv al minții bebelușului. Sistemele superinteligente de inteligență artificială fac față aceleiași provocări când încearcă să înțeleagă comportamentul omenesc.

Probabil va mai trece mult timp până vom avea o inteligență artificială superinteligentă. Pe termen scurt, dezvoltarea unor modele mai bune pentru oameni se poate dovedi extrem de valoroasă pentru orice companie care câștigă bani analizând comportamentul omenesc – adică, în prezent, aproape orice companie care face afaceri pe web. În ultimii câțiva ani, tehnologii comerciale importante pentru interpretarea imaginilor și textului au fost create din dezvoltarea unor modele mai bune pentru vedere și limbaj. Dezvoltarea unor modele bune pentru oameni este următoarea bornă importantă.

Bineînțeles, dacă înțelegem cum funcționează mintea omenească, îmbunătățirea interacțiunii oamenilor cu computerele nu este singurul avantaj. Compromisul dintre a face greșeli și a gândi prea mult ce caracterizează cogniția omenească este un compromis pe care trebuie să-l facă toți agenții inteligenți din lumea reală. Oamenii sunt un exemplu uimitor de sisteme care acționează inteligent în ciuda unor restricții computaționale semnificative. Suntem foarte buni când trebuie să dezvoltăm strategii care ne permit să rezolvăm probleme fără a munci prea mult. Când vom înțelege cum facem acest lucru, vom reuși să facem computerele să lucreze mai inteligent, nu mai mult.

Capitolul 13
Integrarea omului în ecuația inteligenței artificiale

Anca Drăgan

Anca Drăgan este profesor asistent la Departamentul de Inginerie Electrică și Informatică de la UC Berkeley. A coînființat și se află în comitetul de conducere al Laboratorului Berkeley AI Research (BAIR) și este coinvestigator principal la Centrul de Inteligență Artificială Compatibilă cu Omul, de la Universitatea Berkeley.

Cercetarea româncei Anca Drăgan se concentrează pe algoritmi care vor permite roboților să lucreze cu, în jurul și în sprijinul oamenilor. Conduce Laboratorul InterACT de la Berkeley, unde studenții săi lucrează la mai multe aplicații, de la roboți ce oferă asistență la fabricație până la mașini autonome și își bazează activitatea pe controlul optim, planificare, estimare, învățare și științe cognitive. Având treizeci și ceva de ani, a scris deja în colaborare câteva lucrări cu colegul veteran de la Berkeley și mentorul său Stuart Russell, abordând diverse aspecte ale învățării automate și problemele noduroase ale alinierii valorilor.

Împărtășește preocuparea lui Stuart privind siguranța inteligenței artificiale. „Un risc imediat sunt agenții care produc comportamente nedorite, surprinzătoare", a spus ea unui intervievator de la Institutul Future of Life. „Chiar dacă plănuim să folosim inteligența artificială pentru bine, lucrurile pot lua o turnură greșită, în special pentru că nu ne pricepem să specificăm obiectivele și restricțiile agenților de inteligență artificială. Soluțiile lor sunt adesea ceva ce nu ne convine."

Obiectivul său principal este, așadar, să ajute roboții și pe programatori să depășească multele conflicte ce iau naștere din cauza lipsei de transparență cu privire la intențiile celuilalt. Roboții, spune ea, trebuie să ne pună întrebări. Trebuie să-și pună singuri întrebări în legătură cu sarcinile primite și trebuie să stea pe capul programatorilor umani până când toată lumea este de acord – pentru a evita ceea ce a numit ea eufemistic „efecte secundare nedorite".

La baza inteligenței artificiale se află definiția noastră matematică pentru ce înseamnă un agent al inteligenței artificiale (un robot). Când definim un robot, definim stări, acțiuni și recompense. Să ne gândim la un robot care face livrări, de exemplu. Stările sunt locuri din lume, iar acțiunile sunt mișcările pe care robotul le face pentru a ajunge dintr-o poziție în alta apropiată. Pentru a permite robotului să decidă ce acțiune urmează să facă, definim o funcție de recompensare – o mapare a stărilor și acțiunilor în punctaje care arată cât de bună a fost acțiunea respectivă în cadrul stării – și facem robotul să aleagă acțiunile care ar acumula maximul de „recompensă". Robotul obține o recompensă mare când ajunge la destinație și implică un mic cost de fiecare dată când se mișcă; această funcție de recompensare stimulează robotul să ajungă la destinație cât se poate de repede. Similar, o mașină autonomă ar putea obține o recompensă făcând progrese pe ruta sa și ar putea implica un cost dacă se apropie prea mult de alte mașini.

Date fiind aceste definiții, sarcina unui robot este să-și dea seama ce acțiuni trebuie să întreprindă pentru a obține cea mai mare recompensă cumulată. Am lucrat din greu în domeniul inteligenței artificiale pentru a permite roboților să facă acest lucru. Implicit, am presupus că, dacă vom reuși – dacă roboții ar putea lua definiția unei probleme și ar putea-o transforma într-o lege în baza căreia să acționeze –, vom obține roboți care sunt utili oamenilor și societății.

Până acum, nu ne-am înșelat prea mult. Dacă vrei o inteligență artificială care să clasifice celulele drept canceroase sau benigne sau un robot care aspiră covorul din sufragerie când ești la serviciu, avem ce să-ți oferim. Unele probleme reale pot fi definite într-adevăr în izolare, cu stări, acțiuni și recompense clare. Dar cu capacitățile crescânde ale inteligenței artificiale, problemele pe care vrem să le abordăm nu intră atât de ușor în acest cadru. Nu mai putem tăia o bucățică minusculă din lume, după care să o așezăm frumos într-o cutie și să i-o oferim unui robot. A-i ajuta pe oameni începe să însemne a lucra în lumea reală, unde trebuie să interacționeze fizic cu oamenii și să tragă concluzii legate de aceștia. „Oamenii" vor trebui să intre formal undeva în definiția problemei inteligenței artificiale.

Mașinile autonome au fost dezvoltate deja. Vor trebui să împartă drumul cu vehicule care sunt conduse de oameni și cu pietoni și vor trebui să învețe să facă astfel compromisuri între a ne aduce acasă cât se poate de repede și a respecta nevoile celorlalți șoferi. Asistenții personali vor trebui să-și dea seama când avem nevoie de ajutor și în ce măsură și ce sarcini preferăm să ducem la bun sfârșit noi înșine versus ce am prefera să facă ei. Un sistem de asistență pentru decizii sau un sistem de diagnosticare medicală va trebui să ne explice recomandările pentru a le putea înțelege și verifica. Meditatorii automați vor trebui să stabilească ce exemple sunt informative sau ilustrative – nu pentru alte mașini, ci pentru noi, oamenii.

Privind și mai departe în viitor, dacă vrem ca inteligențele artificiale foarte capabile să fie compatibile cu oamenii, nu le putem crea izolate de oameni, după care să încercăm să le facem compatibile; mai degrabă, va trebui să definim inteligența artificială „compatibilă cu omul" încă de la început. Oamenii nu pot fi o idee care vine mai târziu.

Când vine vorba despre roboți reali care ajută oameni reali, definiția standard a inteligenței artificiale nu ne ajută din două motive fundamentale: primul, optimizarea funcției de recompensă a robotului în izolare este diferită de optimizarea ei când robotul

acționează pe lângă oameni, pentru că și oamenii acționează. Luăm decizii în baza propriilor noastre interese, și aceste decizii dictează ce acțiuni putem executa. În plus, raționăm cu privire la robot, adică răspundem la ceea ce credem că face sau va face și la capacitățile pe care credem noi că le are. Acțiunile pe care le face robotul trebuie să se integreze bine cu ale noastre. Aceasta este *problema coordonării.*

În al doilea rând, omul decide în cele din urmă care va fi funcția de recompensă a robotului. Și recompensa are ca scop stimularea comportamentului robotului care corespunde cu ce dorește clientul final, designerul sau chiar societatea întreagă. Consider că roboții capabili, care trec dincolo de sarcinile restrânse și definite clar, vor trebui să înțeleagă acest lucru pentru a obține compatibilitatea cu oamenii. Aceasta este problema *alinierii valorilor.*

Problema coordonării: oamenii nu sunt simple obiecte în mediul înconjurător

Când proiectăm roboți pentru o anumită sarcină, este tentant să-i abstractizăm pe oameni până la uitare. Un asistent personal robotic, de exemplu, trebuie să știe cum să se miște pentru a ridica obiecte, așa că definim această problemă în izolare față de oamenii pentru care robotul urmează să ridice obiecte. Totuși, în timp ce robotul se mișcă, nu vrem să se izbească de lucruri sau de oameni, așa că am putea include localizarea fizică a persoanei în definiția stării robotului. La fel și pentru mașini: nu vrem să se ciocnească cu alte mașini, așa că le permitem să urmărească pozițiile celorlalte mașini și să presupună că se vor deplasa uniform, în aceeași direcție, pe viitor. O ființă umană, în acest sens, nu este diferită pentru un robot față de o minge care se rostogolește pe o suprafață plată. Mingea se va comporta în următoarele câteva secunde în același fel în care s-a comportat și în ultimele secunde, continuă să se rostogolească în aceeași direcție cu aproximativ aceeași viteză. Bineînțeles, aceasta nu seamănă cu comportamentul oamenilor, dar o astfel de simplificare permite multor roboți să-și ducă

sarcinile la bun sfârșit și, în cea mai mare parte, să nu-i încurce pe oameni. Un robot din casa ta, de exemplu, te poate vedea venind pe hol, se poate feri din calea ta pentru a te lăsa să treci și își poate relua sarcina de unde a rămas după ce treci.

Pe măsură ce roboții au devenit din ce în ce mai capabili însă, tratarea oamenilor ca obiecte în continuă mișcare începe să nu mai fie eficientă. Un șofer uman care schimbă banda nu va continua în aceeași direcție, dar va continua să meargă în față după ce a schimbat banda. Când încerci să ajungi la un obiect, adesea te întinzi după el printre alte obiecte și te oprești când ajungi la cel pe care-l dorești. Când mergi pe un hol, ai o destinație în minte: poți face dreapta și intra în dormitor sau poți face stânga și intra în sufragerie. Presupunerea că nu suntem diferiți de o minge care se rostogolește conduce la ineficiență atunci când robotul se ferește dacă nu este necesar și poate pune robotul în pericol când comportamentul persoanei se schimbă. Doar și pentru a nu-i încurca pe oameni, roboții trebuie să fie oarecum precișii în anticiparea acțiunilor umane. Și, spre deosebire de mingea care se rostogolește, ce fac oamenii depinde de ceea ce *decid* să facă. Prin urmare, pentru a anticipa acțiunile omenești, roboții trebuie să înceapă să înțeleagă procesul de luare a deciziilor. Și asta nu înseamnă să presupună despre comportamentul uman că este în întregime optim; poate ar fi suficient pentru un robot care joacă șah sau Go, dar, în lumea reală, deciziile oamenilor sunt mai puțin predictibile decât mutarea optimă pe o tablă de joc.

Această nevoie de a înțelege acțiunile și deciziile umane se aplică atât roboților fizici, cât și celor nefizici. Dacă oricare dintre cele două tipuri își bazează deciziile despre cum să acționeze în legătură cu presupunerea că un om poate face un lucru și face altul, neconcordanța rezultată poate fi catastrofală. Pentru mașini, poate însemna coliziune. Pentru o inteligență artificială cu un rol economic, de exemplu, neconcordanța între ce se așteaptă să facem și ce facem efectiv poate avea consecințe și mai rele.

O alternativă este ca robotul să nu prezică acțiunile umane și să se protejeze, în schimb, de *cel mai rău caz posibil* creat de o

acțiune omenească. Adesea, când roboții fac asta însă, nu mai sunt utili. Cu mașinile, acest rezultat este blocat, pentru că face totul prea riscant.

Toate acestea ne pun pe noi, comunitatea cercetătorilor inteligenței artificiale, într-o situație delicată. Sugerează că roboții vor avea nevoie de modele predictive precise (sau, cel puțin, rezonabile) pentru orice vor decide oamenii să facă. Modul în care încercăm să definim stările nu poate include poziția fizică a oamenilor în lume. În schimb, trebuie să estimăm ceva *intern* pentru oameni. Va trebui să proiectăm roboți ce țin cont de această stare internă a oamenilor, și asta nu va fi ușor. Din fericire, oamenii tind să ofere indicii roboților cu privire la starea lor internă: acțiunile lor în desfășurare oferă robotului observații (în sensul deducțiilor beyesiene) cu privire la intențiile lor. Dacă începem să mergem pe partea dreaptă a holului, probabil vom intra în următoarea încăpere de pe partea dreaptă.

Ce face această problemă și mai complicată este faptul că oamenii nu iau decizii în izolare. Ar fi ceva dacă roboții ar putea preconiza acțiunile pe care urmează să le facă o persoană și să-și dea seama doar ce au de făcut ca răspuns. Dar, din păcate, aceasta ar conduce la roboți ultradefensivi care ar bulversa complet oamenii. (Gândește-te la șoferii umani care se blochează în intersecții cu patru sensuri de mers, de exemplu.) Ce-i lipsește abordării bazate pe preconizarea intenției este că, în momentul în care robotul acționează, acea acțiune a sa influențează acțiunile omului.

Există o influență reciprocă între roboți și oameni, una prin care roboții vor trebui să învețe să navigheze. Problema nu este întotdeauna planificarea robotului în jurul oamenilor; și oamenii planifică în jurul robotului. Este important ca roboții să țină cont de acest lucru când decid acțiunile lor ulterioare, indiferent dacă se află pe drum, în bucătărie sau chiar în spații virtuale, unde o acțiune poate însemna o achiziție sau adoptarea unei noi strategii. Acest lucru ar conferi roboților strategii *de coordonare*, permițându-le să ia parte la negocierile pe care oamenii le fac fără efort în fiecare zi – de la persoana care are prioritate la intersecție sau

printr-o ușă îngustă la rolul pe care ni-l asumăm când colaborăm la pregătirea micului dejun până la obținerea unui consens cu privire la pasul următor dintr-un proiect.

În cele din urmă, la fel cum roboții trebuie să anticipeze ce vor face oamenii, și oamenii trebuie să procedeze la fel cu roboții. Acesta este motivul pentru care transparența este importantă. Roboții nu vor avea nevoie numai de modele mentale bune pentru oameni, dar și oamenii vor avea nevoie de modele mentale bune pentru roboți. Modelul pe care-l are o persoană pentru un robot trebuie să fie integrat în definiția noastră pentru stare, iar robotul trebuie să fie conștient de modul în care acțiunile sale modifică modelul respectiv. Asemănător robotului care tratează acțiunile umane ca indicii pentru stările interne ale omului, oamenii își vor schimba părerile despre robot pe măsură ce-i observă acțiunile. Din păcate, roboții nu dau indicii la fel de ușor ca oamenii; noi am exersat mult să comunicăm implicit cu oamenii. Dar a permite roboților să conștientizeze modificarea pe care o cauzează acțiunile lor în modelul mental al persoanei pentru robot poate conduce la acțiuni mai atent alese care conferă indicii corecte – care comunică în mod clar oamenilor intențiile robotului, funcția sa de recompensă și limitările sale. De exemplu, un robot își poate schimba mișcarea când cară ceva greu pentru a sugera dificultatea pe care o are în manevrarea obiectelor grele. Cu cât oamenii știu mai multe despre robot, cu atât este mai ușor să se coordoneze cu el.

Obținerea compatibilității în acțiuni va necesita ca roboții să anticipeze acțiunile omenești, să țină cont de modul în care acele acțiuni le vor influența pe ale lor și să permită oamenilor să anticipeze acțiunile roboților. Cercetarea a făcut progrese în vederea depășirii acestor obstacole, dar mai avem un drum lung de parcurs.

Problema alinierii valorii: oamenii dețin cheia funcției de recompensă a roboților

Progresul care va permite roboților să optimizeze recompensa pune o povară mai grea pe umerii noștri, ai designerilor, pentru a le conferi recompensa potrivită care trebuie optimizată în primul rând. Ideea inițială era că, pentru orice sarcină pe care voiam să i-o dăm robotului, am putea scrie o funcție de recompensare ce oferă stimulente pentru comportamentul potrivit. Din păcate, ce se întâmplă adesea este că specificăm o anumită funcție de recompensare și comportamentul care reiese din optimizarea ei nu este ceea ce vrem. Funcțiile de recompensare intuitivă, atunci când se combină cu instanțe neobișnuite ale unei sarcini, pot conduce la comportamente neintuitive. Recompensezi un agent dintr-un joc de curse cu un punctaj în joc și, în unele cazuri, găsește o portiță pe care o exploatează pentru a câștiga mult mai multe puncte fără a câștiga efectiv cursa. Stuart Russell și Peter Norvig oferă un exemplu minunat în cartea lor *Artificial Intelligence: A Modern Approach*: recompensarea unui robot de aspirat pentru cât de mult praf aspiră va face ca acesta să scoată din praful deja aspirat și să-l aspire din nou pentru a obține o recompensă mai mare.

În general, oamenilor le-a fost foarte greu să specifice exact ce vor, după cum arată toate acele povești cu djinni. O paradigmă a inteligenței artificiale în care roboții pot obține recompense specificate extern eșuează atunci când recompensa nu este gândită perfect. Poate oferi stimulente robotului să se comporte greșit și chiar să reziste încercărilor noastre de a-i corecta comportamentul, pentru că această corecție ar conduce la o recompensă mai mică.

O paradigmă aparent mai bună ar fi ca roboții să optimizeze ceea ce ne dorim intern, chiar dacă ne este mai greu să explicăm. Aceștia ar urma să folosească ce spunem și ce facem ca dovezi pentru ceea ce vrem mai degrabă decât a interpreta asta literal sau ca pe un dat. Când scriem o funcție de recompensare, robotul ar trebui să înțeleagă că ne putem înșela; că este posibil să nu fi luat

în considerare toate fațetele sarcinii; că nu există nicio garanție că funcția de recompensare dată va conduce *întotdeauna* la comportamentul pe care ni-l dorim. Robotul ar trebui să integreze ce am scris noi în înțelegerea sa cu privire la ce ne dorim noi, dar ar trebui și să distileze informații clarificatoare din comportamentul nostru. Ar trebui să urmărească ghidarea noastră, pentru că acesta este singurul mod de a optimiza funcția de recompensare dorită cu adevărat.

Chiar dacă reușim să conferim roboților capacitatea de a învăța ce dorim, rămâne o întrebare importantă la care inteligența artificială nu poate răspunde singură. Putem face roboții să încerce să-și alinieze valorile cu valorile interne ale unei persoane, dar problema este că e vorba despre mai multe persoane. Robotul are un utilizator final (sau, poate, câțiva, cum ar fi un robot personal care îngrijește o familie, o mașină care conduce mai mulți pasageri în mai multe destinații sau un asistent de birou pentru o echipă întreagă); are un designer (sau, poate, câțiva); și interacționează cu societatea – mașina autonomă împarte drumul cu pietoni, vehicule conduse de oameni și alte mașini autonome. Cum vom combina valorile acestor entități când ar putea intra în conflict este o problemă importantă pe care trebuie să o rezolvăm. Cercetarea din domeniul inteligenței artificiale ne poate oferi instrumentele necesare pentru a combina valorile în orice fel decidem, dar nu poate lua deciziile necesare pentru noi.

Pe scurt, trebuie să permitem roboților să emită judecăți în privința noastră – să ne vadă ca pe ceva mai mult decât simple obstacole sau jucători perfecți. Trebuie să ne ia natura umană în considerare, astfel încât să fie bine coordonați și bine aliniați cu noi. Dacă reușim, vom avea într-adevăr instrumente care ne vor spori substanțial calitatea vieții.

Capitolul 14
Coborârea pe panta unui gradient

Chris Anderson

Chris Anderson este antreprenor; fost redactor-șef al revistei *Wired*, cofondator și CEO al companiei 3DR și autorul cărților *The Long Tail, Free* și *Makers*.

Compania lui Chris Anderson, 3DR, a ajutat la lansarea industriei moderne a dronelor și se concentrează acum pe softul datelor pentru drone. S-a lansat construind o comunitate de robotică aeriană cu sursă deschisă denumită DIY Drones* și a făcut câteva experimente timpurii nerecomandate, cum ar fi să bâzâie la Laboratorul Lawrence Berkeley cu unul dintre spionii săi zburători autonomi. Putea foarte bine să fie un caz de expresie a genelor antice, de vreme ce este descendentul unui fondator al mișcării anarhiste din America. Chris a condus revista *Wired*, publicația preferată a tehno-utopiștilor, dar și a tehno-distopiștilor, din 2001 până în 2012; în timpul mandatului său, revista a câștigat cinci premii National Magazine.

* DIY, acronim pentru Do It Yourself folosit des în limba engleză, ce exprimă ideea de a face ceva pe cont propriu, fără ajutorul profesioniștilor (n.t.).

Chris nu agreează termenul „robotician" („ca orice robotician modest, eu nu mă consider unul"). A început ca fizician. „M-am dovedit a fi un fizician slab", mi-a spus el recent. „M-am luptat, am mers la Los Alamos și m-am gândit «Ei bine, poate că n-o să fiu laureat al Premiului Nobel, dar pot să fiu totuși un om de știință». Toți cei care eram în domeniul fizicii și aveam acești eroi romantici – frații Feynman, proiectul Manhattan – ne-am dat seama că traiectoria carierei noastre ne-ar duce în cel mai bun caz la un proiect de la CERN în cincisprezece ani. Acel proiect urma să fie un eșec, caz în care nu ar apărea nicio lucrare, sau un succes, caz în care ai fi autorul numărul trei sute pe lucrarea respectivă și ai deveni profesor asistent la Universitatea de Stat din Iowa."

„Majoritatea colegilor mei au mers pe Wall Street și au devenit quanți*, și acestora le datorăm criza imobiliarelor. Alții au pus bazele internetului. Mai întâi, am construit internetul conectând laboratoarele de fizică. Apoi am construit webul. Apoi am fost primii care am folosit Big Data. Aveam supercomputere – Crays – care aveau jumătate din puterea telefoanelor noastre de acum, dar erau supercomputerele vremii. Între timp, citeam această revistă numită *Wired*, care a apărut pe piață în 1993, și ne-am dat seama că acest instrument pe care-l folosim noi, oamenii de știință, ar putea fi folosit de toată lumea. Internetul nu fusese creat doar pentru date științifice, era o revoluție culturală uluitoare. Așa că, atunci când Condé

* De la Quant, acronim pentru Quantitative Analysis for Economics and Finance – analiză cantitativă pentru economie și finanțe. Termen folosit și cu sensul de investitori specializați de pe Wall Street care-i ajută pe moguli să câștige și mai mulți bani (n.t.).

Nast mi-a cerut să preiau conducerea revistei, am spus «Clar!». Această revistă mi-a schimbat viața."

Avea cinci copii la vremea aceea – toți jucau jocuri video – care l-au făcut să fie interesat de „roboții zburători". Și-a dat demisia de la *Wired*. Restul este o istorie din Silicon Valley.

Viața

Țânțarul îmi detectează mirosul de la zece metri. Mirosul meu îi declanșează funcția de urmărire, care constă în cele mai simple reguli posibile. Mai întâi, deplasarea într-o direcție aleatorie. Dacă mirosul se intensifică, continuă să se deplaseze în direcția respectivă. Dacă mirosul scade, se deplasează în partea opusă. Dacă mirosul dispare, se mișcă în lateral până când îl regăsește. Repetă acest set până la atingerea țintei.

Norul mirosului meu este cel mai dens lângă mine și dispare pe măsură ce distanța crește, o ceață invizibilă de particule emanată din pielea mea care se mișcă precum fumul în vânt. Cu cât este mai aproape de pielea mea, cu atât este mai mare densitatea particulelor; cu cât este mai departe, cu atât este mai mică. Scăderea este numită gradient, cuvânt ce descrie o tranziție progresivă de la un nivel la altul – în opoziție cu funcția continuă, care descrie o schimbare discretă.

Când țânțarul urmărește acest gradient până la sursă folosindu-și algoritmul simplu, aterizează pe pielea mea, pe care o simte cu senzorii termici din picioare, care sunt acordați la alt gradient: temperatura. Apoi își împinge trompa sub formă de ac prin suprafață, unde al treilea set de senzori din vârful acesteia detectează alt gradient, acela al densității sangvine. Acul flexibil se mișcă pe sub pielea mea până când mirosul de sânge îl ghidează

către un vas capilar, pe care îl perforează. Apoi sângele meu începe să curgă în țânțar. Misiune îndeplinită. Au!

Ce pare un radar puternic al insectelor în întuneric, cu o inteligență inexplicabilă pentru niște creiere atât de mici în ceea ce privește căutarea sângelui, este, de fapt, doar un nas sensibil, lipsit aproape complet de inteligență. Țânțarii sunt mai apropiați de plantele care caută soarele decât de rachetele ghidate. Însă, dacă aplicăm literal această regulă „urmează-ți nasul", se pot deplasa prin toată casa și te pot găsi, se pot strecura prin crăpăturile din ușa cu plasă și pot chiar identifica bucățica de piele pe care ai lăsat-o neacoperită dintre pălărie și gulerul cămășii. Este doar un mers aleatoriu, combinat cu aripi flexibile și picioare care permit insectelor să ricoșeze din obstacole, precum și un instinct de a coborî pe panta unui gradient chimic.

Dar „coborârea pe panta unui gradient" este mult mai mult decât metoda de navigare a insectelor. Privește în jurul tău și o vei vedea peste tot, de la cele mai simple reguli fizice ale universului până la cea mai avansată inteligență artificială.

Universul

Trăim într-o lume compusă din nenumărați gradienți, de la lumină și căldură la gravitație și dâre chimice (dâre chimice!*) Apa curge de-a lungul unui gradient al gravitației, iar corpul tău trăiește din soluții chimice care alunecă pe membranele celulare de la o concentrație mare înspre una mică. Toate acțiunile din univers au la bază un motor gradient, de la mișcarea planetelor de-a lungul gradienților gravitaționali până la fuziunea atomilor de-a lungul gradienților cu încărcătură electrică pentru a forma molecule. Propriile noastre trebuințe, cum ar fi foamea și somnul, au ca motor gradienți electrochimici aflați în corpurile noastre. Și

* Teorie conspiraționistă și pseudoștiințifică, respinsă de majoritatea oamenilor de știință (n.t.).

funcțiile creierului nostru, semnale electrice care se deplasează pe canalele ionizate din sinapsele ce ne conectează neuronii, sunt pur și simplu atomi și electroni care curg „în aval" pe alți gradienți electrici și chimici. Uită de analogiile cu ceasurile, creierele noastre sunt mai aproape de un sistem de canale și încuietori, cu semnale care se deplasează ca apa dintr-o stare în alta.

În timp ce stau aici și scriu aceste rânduri, caut, de fapt, stări de echilibru într-o topologie *n*-dimensională de gradienți. Să ne gândim doar la unul: căldura. Temperatura corpului meu este mai ridicată decât temperatura ambientală, așa că radiez căldură, care trebuie înlocuită în corpul meu. Chiar și bacteriile din tractul meu digestiv folosesc senzori care măsoară concentrațiile de zahăr din lichidul ce le înconjoară și dau din cilii lor asemănători unei cozi pentru a înota „în amonte", unde depozitele de zahăr sunt cele mai bogate. Starea naturală a tuturor sistemelor este să curgă către stările cu energie mai scăzută, un proces care este descris în sens larg ca entropie (tendința lucrurilor de a aluneca din stări de organizare în stări de dezorganizare; totul se va destrăma în cele din urmă, inclusiv Universul însuși).

Dar cum explicăm comportamentele complexe, cum ar fi capacitatea noastră de a lua decizii? Răspunsul este în continuare coborârea pe panta unui gradient.

Creierele noastre

Pe cât este de miraculoasă și enigmatică inteligența omenească, știința începe să accepte ideea despre creierele noastre că ar funcționa în același fel cu alte sisteme complexe cu straturi și cicluri de *feedback*, toate urmărind ceea ce numim matematic „funcții de optimizare", dar am putea la fel de bine să le numim „curgere în aval" într-un anumit sens.

Esența inteligenței este învățarea și învățăm corelând informațiile de intrare cu punctaje pozitive sau negative (recompense sau pedepse). Prin urmare, pentru un bebeluș, „acest sunet" (vocea mamei tale) este asociat cu alte conexiuni învățate cu „mama ta",

cum ar fi mâncarea și confortul. Asemenea „această mișcare a mușchiului îmi aduce degetul mai aproape de gură". În timp, prin încercare și eroare, rețeaua neuronală a creierului întărește aceste conexiuni. Între timp, „această mișcare a mușchiului nu îmi aduce degetul mai aproape de gură" este o corelație negativă, iar creierul va slăbi acele conexiuni.

Însă acest model este prea simplist. Limitele coborârii pe panta unui gradient constituie așa-numita problemă „minima locală" (sau problema „maxima locală" dacă avem de-a face cu *urcarea* pe panta unui gradient). Dacă mergi printr-o regiune muntoasă și vrei să ajungi acasă și dacă mergi tot timpul în jos, vei ajunge cu siguranță în următoarea vale, dar nu neapărat și peste ceilalți munți care sunt în jurul ei și care se află între tine și casă. Pentru a ajunge la destinație, ai nevoie de un model mental (de exemplu, o hartă) al topologiei, pentru a ști unde trebuie să urci și să ieși de pe vale, sau trebuie să alternezi între coborâri pe panta gradientului și plimbări aleatorii pentru a putea ieși din regiune.

Asta face, de fapt, țânțarul când urmărește mirosul meu. Coboară când se află în norul meu și se deplasează aleatoriu când pierde urma sau dă de un obstacol.

Inteligența artificială

Așadar, asta-i natura. Dar computerele? Softurile tradiționale nu funcționează așa – respectă arbori determiniști de logică dură: „dacă așa, atunci fă așa". Dar softurile care interacționează cu lumea fizică au tendința să funcționeze *asemănător* lumii fizice. Asta înseamnă că au de-a face cu date de intrare zgomotoase (senzori sau comportament uman) și oferă rezultate probabiliste, și nu deterministe. Și acest lucru, la rândul său, înseamnă și mai mult o coborâre pe panta unui gradient.

Softurile inteligenței artificiale sunt cele mai bune exemple, în special tipurile de inteligență artificială care utilizează modele artificiale de rețele neuronale (inclusiv rețele neuronale spiralate

sau „profunde", cu multe straturi). În acestea, un proces tipic constă în „instruirea" lor, arătându-le multe exemple pentru ce ne dorim să învețe (imagini cu pisici care au eticheta „pisică" de exemplu), alături de exemple cu alte date aleatorii (imagini cu altceva). Această metodă este denumită „învățare supravegheată", deoarece rețeaua neuronală este educată prin exemple, inclusiv utilizarea „instruirii adversative", cu date care nu sunt corelate cu rezultatele dorite.

Aceste rețele neuronale, la fel ca modelele lor biologice, constau în straturi cu mii de noduri („neuroni", în analogie), fiecare fiind conectat cu toate nodurile din straturile inferioare și superioare prin conexiuni care au la început o forță aleatorie. Nivelul superior primește date și nivelul inferior generează răspunsul corect. Orice serie de conexiuni care s-a întâmplat să aterizeze pe răspunsul corect este întărită („recompensată"), iar cele care s-au înșelat sunt slăbite („pedepsite"). Dacă se repetă acest proces de zeci de mii de ori, avem de-a face cu o rețea complet instruită pentru acest tip de date.

Ne putem gândi la toate combinațiile posibile de conexiuni ca la suprafața unei planete, cu dealuri și văi. (Ignorăm momentan că suprafața este doar tridimensională și topologia efectivă este multidimensională.) Optimizarea prin care trece rețeaua în timp ce învață este asemănătoare procesului de a găsi cea mai adâncă vale de pe planetă. Constă în următoarele etape:

1. Definirea unei „funcții de cost" care stabilește cât de bine a rezolvat rețeaua problema.
2. Rularea rețelei o dată și analiza rezultatului generat de funcția de cost respectivă.
3. Modificarea valorilor conexiunilor și repetarea procesului. Diferența dintre cele două rezultate este direcția sau „panta" pe care rețeaua s-a deplasat între cele două încercări.
4. Dacă panta este „în jos", conexiunile sunt modificate mai mult în acea direcție. Dacă este „în sus", sunt modificate în direcția opusă.

5. Repetarea procedurii până când nu mai apare nicio îmbunătățire în nicio direcție. Ceea ce înseamnă că ne aflăm la punctul minim.

Felicitări! Dar, probabil, este un minim *local* sau o mică adâncitură în munte, așa că va trebui să continuăm dacă vrem ca rezultatul să fie și mai bun. Nu mai putem continua în jos și nu știm unde este cel mai jos punct, așa că trebuie să-l găsim cumva. Există mai multe posibilități, dar iată câteva:

1. Încercăm de multe ori, cu setări aleatorii diferite, și împărtășim rezultatele fiecărei încercări; în fond, zgâlțâim sistemul pentru a vedea dacă se liniștește undeva mai jos. Dacă unul dintre celelalte rezultate a găsit o vale mai adâncă, începem de la acele setări.
2. Nu mergem încontinuu în jos, ci și „pe trei cărări", ca un bețiv (metodă denumită „coborâre stocastică pe panta unui gradient"). Dacă facem asta suficient timp, vom găsi în cele din urmă punctul cel mai jos. E ca o metaforă pentru viață.
3. Căutăm trăsături „interesante" definite de diversitate (limite sau schimbări de culoare, de exemplu). Avertizare: această metodă poate duce la nebunie – prea mult „interesant" atrage rețeaua către iluzii optice. Așa că am face bine să rămânem sănătoși și să punem accent pe tipul de trăsături care prezintă o probabilitate mai mare de a fi reale în natură, spre deosebire de artefacte sau erori. Acest proces este numit „regularizare" și există multe tehnici pentru a-l obține, cum ar fi dacă acele caracteristici au mai fost văzute (învățate) sau sunt „prea frecvente" (ca zgomotul de fundal), și nu „mai puțin frecvente" (mai continue, ca trăsăturile reale din natură).

Doar pentru că sistemele de inteligență artificială ajung uneori în valoarea minimă locală, să nu tragem concluzia că asta le face mai diferite de viață. Oamenii – și, probabil, toate formele de viață – sunt adesea blocați în valoarea minimă locală.

Să luăm, de exemplu, înțelegerea noastră în ceea ce privește jocul de Go, care a fost predat, învățat și optimizat de oameni mii de ani. I-a luat inteligenței artificiale trei ani să afle că l-am jucat greșit tot acest timp și că există soluții mai bune, aproape extraterestre, pentru acest joc pe care nu le-am luat niciodată în considerare – în cea mai mare parte, deoarece creierele noastre nu au puterea de procesare pentru a se gândi la atât de multe mutări în avans.

Chiar și în șah, care este de zece ori mai simplu și despre care credeam că l-am înțeles, mașinile cu forță brută ne pot învinge cu propriile noastre strategii. Și șahul, se pare, când este explorat de sisteme de inteligență artificială cu rețele neuronale superioare, are strategii bizare, dar superioare la care nu ne-am gândit, cum ar fi sacrificarea timpurie a reginei pentru a câștiga un avantaj obscur pe termen lung. Este ca și cum am fi jucat versiuni bidimensionale ale unor jocuri care există în dimensiuni superioare.

Dacă ceva din ce spun îți sună cunoscut, este pentru că fizica se luptă cu acest tip de probleme topologice de decenii întregi. Noțiunea de spațiu fiind pluridimensională și matematica redusă la înțelegerea geometriilor și interacțiunilor „membranelor" care se află dincolo de simțurile noastre, este locul în care teoreticienii marii unificări merg să moară. Dar, spre deosebire de fizica teoretică multidimensională, inteligența artificială este ceva ce putem experimenta și măsura.

Prin urmare, asta vom face. Următoarele câteva decenii vor fi o explorare explozivă a modalităților de gândire pe care 7 milioane de ani de evoluție nu le-au găsit. Vom da din coate până ieșim din valoarea minimă locală și vom găsi valori minime mai profunde, poate chiar valoarea minimă globală. Și, după ce le găsim, poate chiar învățăm mașinile să pară la fel de inteligente ca un țânțar, coborând încontinuu pe panta gradienților cosmici până la un ultim scop, oricare ar fi acesta.

Capitolul 15
„Informația" pentru Wiener, pentru Shannon și pentru noi

David Kaiser

David Kaiser este profesor Germeshausen de istorie a științei și profesor de fizică la MIT, director al programelor sale științifice, tehnologice și sociale. Este autorul cărților *How the Hippies Saved Physics: Science, Counterculture, and the Quantum Revival* și *American Physics and the Cold War Bubble* (în curs de apariție).

David Kaiser este un fizician atipic, interesat de intersecția științei sale cu politica și cultura, despre care a scris mult.

La prima întâlnire (din Washington, Connecticut), care a precedat apariția acestei cărți, a comentat despre modul în care s-a modificat perspectiva asupra „informației" de pe vremea lui Wiener: era complexului militar industrial din Războiul Rece. Pe atunci, Wiener compara informația, la nivel metaforic, cu entropia, în sensul că nu poate fi conservată, adică monopolizată; astfel, argumenta el, secretele noastre atomice și alte probleme clasificate nu ar rămâne secrete prea mult timp. Astăzi, întrucât informația (așa cum s-ar fi așteptat Wiener), falsă sau nu, se scurge peste tot prin celălalt Washington, informațiile din lumea

economică au fost într-adevăr colectate, comunicate și monetizate.

Această fixare, spune David, „nu era pe de-a-ntregul bună și nici pe de-a-ntregul rea" – dacă, bănuiesc, te-ai săturat să fii sâcâit de reclame la șosete sau croaziere pe râuri europene în browserul tău la un minut după ce le-ai cumpărat deja.

Ca să nu spunem nimic despre proliferarea informației. David ni s-a plâns nouă, participanților la întâlnire, că, pe vremea lui Wiener, fizicienii puteau „citi toată revista *Physical Review*. Stătea confortabil în fața noastră într-o grămadă pe care o puteam gestiona. Acum suntem potopiți cu cincizeci de mii de reviste cu sursă deschisă pe minut", pline cu Dumnezeu știe ce. Wiener nu a anticipat niciuna dintre aceste dezvoltări, spune David, făcându-l să întrebe „Avem nevoie oare de un set nou de metafore după care să ne ghidăm?"

În *The Sleepwalkers*, o istorie cuprinzătoare a gândirii științifice din vremurile antice până în Renaștere, Arthur Koestler a identificat o tensiune care a marcat cele mai dramatice salturi ale imaginației noastre cosmologice. Citind dintr-o perspectivă prezentă lucrările lui Nicolaus Copernicus și Johannes Kepler, Koestler susținea că suntem loviți de nefamiliaritatea lor bizară – integrarea lor în magia sau misticismul unei epoci trecute –, dar și de observațiile lor care par a fi moderne.

Eu văd aceleași sensuri duble – zigzagul de pliuri origami de vechi și nou – în bine-cunoscuta carte a lui Norbert Wiener *The Human Use of Human Beings*. Publicată inițial în 1950 și revizuită în 1954, cartea este în multe sensuri extraordinar de profetică. Wiener, eruditul de la MIT, a recunoscut în fața majorității observatorilor că „societatea poate fi înțeleasă doar prin studierea mesajelor și instalațiilor de comunicare ce îi aparțin". Wiener susținea că ciclurile de *feedback*, elementul central din întreaga sa teorie cibernetică, vor urma să joace un rol central în dinamica socială. Aceste cicluri nu doar că-i vor conecta pe oameni între ei, dar îi vor conecta și pe oameni cu mașinile și, crucial, mașinile cu mașinile.

Wiener a întrezărit o lume în care informația va putea fi separată de mediul său. Oamenii sau mașinile ar putea comunica tiparele lor pe distanțe uriașe și le vor putea utiliza pentru a produce obiecte noi în punctele de destinație, fără să „miște o... particulă de materie de la un capăt al liniei la celălalt", o viziune realizată acum în lumea noastră de imprimante 3D în rețea.

Wiener și-a imaginat, de asemenea, și cicluri de *feedback* de la o mașină la alta, care vor sta la baza unor avansuri uriașe în automatizare, chiar și pentru sarcini care s-au bazat înainte pe judecata umană. „Mașina nu are preferințe când trebuie să aleagă între munca fizică și munca de birou", a observat el.

Cu toate acestea, multe dintre argumentele centrale din *The Human Use of Human Beings* par să fie mai apropiate de secolul al XIX-lea decât de secolul al XXI-lea. În special, deși Wiener a făcut referire în toată lucrarea sa la ceea ce era pe atunci noua teorie a informației emisă de Claude Shannon, se pare că nu a îmbrățișat cu totul noțiunea lui Shannon de informație care constă în biți ireductibili și fără înțeles propriu. Din vremea lui Wiener, teoria lui Shannon a stat la baza avansurilor recente în domeniile „Big Data" și „învățare automată", ceea ce face cu atât mai interesantă revizitarea imaginației cibernetice a lui Wiener. Cum ar putea fi diferită inteligența artificială de mâine dacă practicanții ar fi să reinvestească în viziunea călăuzitoare a lui Wiener despre „informație"?

Când Wiener scria *The Human Use of Human Beings*, experiențele sale în cercetările legate de război și de ceea ce i se păreau ambiguitățile morale ale vieții intelectuale în cadrul complexului militar industrial erau încă proaspete. Cu doar câțiva ani înainte, anunțase în paginile revistei *Atlantic Monthly* că nu va „publica lucrări viitoare care ar putea cauza rău în mâinile militariștilor iresponsabili"*. A rămas ambivalent cu privire la puterea transformatoare a noilor tehnologii, nelăsându-se nici dus de valul tendințelor la modă și nici de utopismul digital al savanților care i-au urmat.

„Progresul impune nu doar posibilități noi pentru viitor, dar și restricții noi", scria el în *The Human Use of Human Beings*. Era îngrijorat atât de restricțiile umane, cât și de cele tehnologice, în

* Norbert Wiener, „A Scientist Rebels", Atlantic Monthly, ianuarie 1947 (n.a.).

special restricțiile din timpul Războiului Rece care amenințau fluxul de informații atât de important pentru sistemele cibernetice: „Sub elanul senatorului [Joseph] McCarthy și al imitatorilor săi, clasificarea oarbă și excesivă a informațiilor militare" îi determina pe liderii politici din Statele Unite să adopte o „mentalitate atât de discretă încât era egalată doar de perioada Veneției renascentiste". Wiener, spunând ce au spus și mulți veterani ai Proiectului Manhattan, argumenta că obsesia confidențialității după război – în special în privința armelor nucleare – își avea rădăcinile în neînțelegerea procesului științific. Singurul secret veritabil în ceea ce privește producția armelor nucleare, scria el, era dacă astfel de bombe pot fi construite cu adevărat. După dezvăluirea acestui secret, cu bombardarea orașelor Hiroșima și Nagasaki, oricât de mult ar încerca statul să acopere secretul, oamenii nu ar putea fi opriți să gândească etapele urmate de acei cercetători din Proiectul Manhattan. În memorabilele cuvinte ale lui Wiener, „nu există linia Maginot pentru creier".

Pentru a-și demonstra spusele, Wiener a împrumutat ideile proaspete ale lui Shannon despre teoria informației. În 1948, Shannon, matematician și inginer care lucra la Bell Labs, a publicat două articole lungi în revista *Bell System Technical Journal*. Prezentând aceste idei noi pentru mai multe categorii de cititori în 1949, matematicianul Warren Weaver a explicat că, în formularea lui Shannon, „cuvântul *informație*... este utilizat cu un sens special care nu trebuie confundat cu utilizarea sa comună. În mod special, *informația* nu trebuie confundată cu sensul". Lingviștii și poeții se preocupă de aspectele „semantice" ale comunicării, continua Weaver, dar nu și inginerii ca Shannon. Mai degrabă, „acest cuvânt, *informație*, în teoriile comunicării se referă nu atât de mult la *ceea ce spui*, ci mai mult la *ceea ce ai putea spune*". În

* Warren Weaver, „Recent Contributions to the Mathematical Theory of Communication", în Claude Shannon și Warren Weaver, *The Mathematical Theory of Communication* (Urbana, IL: University of Illinois Press, 1949), 8 (text evidențiat în original). Lucrările lui Shannon din 1948 au fost republicate în același volum (n.a.).

cuvintele acum renumite ale lui Shannon, conținutul de informație al unui șir de simboluri era dat de logaritmul numărului de simboluri posibile din care a fost ales un șir dat. Observațiile-cheie ale lui Shannon erau că informația unui mesaj era asemănătoare entropiei unui gaz: o măsură a dezordinii din sistem.

Wiener a împrumutat această observație când a scris *The Human Use of Human Beings*. Dacă informația era asemănătoare entropiei, atunci nu putea fi conservată sau controlată. Fizicienii secolului al XIX-lea au dovedit că energia totală a unui sistem fizic trebuie să rămână întotdeauna constantă, un echilibru perfect între începutul și sfârșitul unui proces. Lucrurile nu stau la fel și cu entropia, care ar crește inexorabil în timp, un imperativ care a ajuns să fie cunoscut drept a doua legii a termodinamicii. Din această puternică diferență – energia este conservată, iar entropia trebuie să crească – reieșeau consecințe cosmice enorme. Timpul trebuie să curgă înainte; viitorul nu poate fi identic cu trecutul. Universul se poate îndrepta către o „moarte termică", într-un timp foarte îndepărtat, când stocul de energie va fi dispersat uniform, ajungând într-o stare de entropie maximă, după care nu mai poate apărea nicio schimbare ulterioară.

Dacă informația *ca* entropie *per se* nu ar putea fi conservată, Wiener a conchis că este o nebunie ca liderii militari să încerce să acumuleze „*know-how*-ul științific al națiunii în biblioteci și laboratoare statice". Într-adevăr, „indiferent de volumul de cercetări științifice, înregistrate cu atenție în cărți și lucrări de specialitate și introduse apoi în bibliotecile noastre cu etichete de secretizare, nu ne vom putea proteja într-o lume în care nivelul efectiv de informație este în continuă creștere". Orice astfel de efort de secretizare, clasificare sau colecționare a acestor informații era sortit pieirii, spunea Wiener, la fel cum escrocheriile pungașilor pentru mașini cu mișcare perpetuă nu făceau față celei de-a doua legi a termodinamicii.

Wiener a criticat „ortodoxia" americană a fundamentalismului pieței libere aproape în același fel. Pentru majoritatea americanilor, „problemele privind informația vor fi evaluate conform unui

criteriu american standard: un lucru este valoros ca bun de consum pentru ce introduce pe piața deschisă". Într-adevăr, „soarta informației în lumea tipic americană urma să devină ceva ce poate fi cumpărat sau vândut"; majoritatea oamenilor, observa el, „nu pot concepe ideea ca informația să nu aibă un proprietar". Wiener considera această idee la fel de greșită precum clasificările militare exagerate. Din nou, invoca observațiile lui Shannon: „deoarece informațiile și entropia nu sunt conservate", acestea sunt „la fel de nepotrivite pentru a deveni bunuri de consum".

Informația nu poate fi conservată – până aici, totul este clar. Dar se referea Wiener la „informația" lui Shannon? Partea dificilă din argumentul lui Shannon, după cum observa Wiener, era să distingă sensul colocvial al „informației", ca mesaj cu sens, de noțiunea abstractă, rarefiată de șiruri de simboluri aranjate cu o anumită probabilitate și selectat dintr-un univers enorm de vorbe fără noimă. Pentru Shannon, „informația" putea fi cuantificată, deoarece unitatea sa fundamentală, bitul, era o unitate de transmitere, și nu una de sens.

Când Wiener a caracterizat „informația" în cartea *The Human Use of Human Beings*, pe de altă parte, a făcut referire de mai multe ori la un sens clasic, umanist al termenului. „O bucată de informație", scria el – atenție, nu un „bit" de informație –, „pentru a putea contribui la informația generală a comunității, trebuie să spună ceva substanțial diferit de stocul de informații comun anterior al comunității". Acesta este motivul pentru care „școlarii nu-l apreciază pe Shakespeare", concluziona el: stihurile bardului pot porni brusc din fluxuri aleatorii de biți, dar au devenit cu toate acestea foarte cunoscute publicului decodificator de sens și „absorbite în clișeele superficiale ale vremii".

Cel puțin, conținutul informațional al lui Shakespeare a părut cândva proaspăt. În anii de după *boom*-ul postbelic, Wiener se văita că „volumul enorm de comunicare *per capita*" – de la ziare și filme până la radio, televiziune și cărți – a generat mediocritate, o revenire informațională la medie. „Din ce în ce mai mult, trebuie

să acceptăm un produs standardizat inofensiv și nesemnificativ care, la fel ca pâinea albă de la brutărie, este făcut mai mult pentru proprietățile sale de păstrare și vandabilitate decât pentru valoarea nutritivă". „Să ne ferească Dumnezeu", exclama el, „de primele romane scrise, pentru că un tânăr dorește prestigiul de a fi romancier, și nu că ar avea ceva de spus! Să ne ferească Dumnezeu și de lucrările matematice care sunt corecte și elegante, dar fără trup și spirit". Felul în care privea Wiener „informația" suna mai mult a Matthew Arnold în 1869* decât a Claude Shannon în 1948 – mai mult „trup și spirit" decât „bit". Și Wiener împărtășea viziunea romantică a lui Arnold a „producătorului de conținut": „la drept vorbind, artistul, scriitorul și omul de știință ar trebui să fie mișcat de un impuls atât de irezistibil de a crea încât, chiar dacă nu ar fi plătit pentru munca depusă, să fie dispus să plătească el însuși pentru șansa de a o face". *L'art pour l'art*, acel strigăt al secolului al XIX-lea: artiștii trebuie să sufere pentru munca lor; căutarea expresiei cu sens trebuie să bată întotdeauna câștigul.

Pentru Wiener, aceasta era măsura corectă a „informației": trup, spirit, aspirație, expresie. Însă, pentru a argumenta împotriva transformării acesteia în bun de consum, Wiener s-a întors încă o dată la matematica lui Shannon care privea „informația ca entropie".

Sărim până în zilele noastre. În multe feluri, s-a dovedit că Wiener avea dreptate. Viziunea sa privind ciclurile de *feedback* în rețea guvernate de comunicații de la mașină la mașină a devenit o caracteristică obișnuită a vieții de zi cu zi. În plus, încă din cele mai vechi zile ale Epocii Internetului, pirateria digitală a desființat imaginea conform căreia „informația" – sub formă de melodii, filme, cărți sau cod – poate rămâne închisă. Pune un abonament cu *paywall* aici și conținutul va fi difuzat acolo, atât de multă entropie informațională încât nu poate fi conservată.

---

* Matthew Arnold, *Culture and Anarchy*, ed. Jane Garnett (Oxford, UK: Oxford University Press, 2006) (n.a.).

Pe de altă parte, corporații multinaționale enorme – câteva dintre cele mai mari și mai profitabile din lume – dezaprobă acum la ordinea zilei punctele de vedere controversate care susțin că „informația" nu poate fi stocată sau monetizată. Ironic, „informația" pe care o comercializează ele este mai aproape de definiția lui Shannon decât de cea a lui Wiener, neținând cont de dovezile matematice ale lui Shannon.

Deși Google Books ajută sute de mii de cărți de literatură să circule gratuit, Google în sine – la fel ca Facebook, Amazon, Twitter și multe alte companii care le imită – a confiscat o formă mai simplă de „informație" și a exploatat-o pentru un profit extraordinar. *Petabytes* de informații de tip Shannon – un șir aparent fără valoare de clicuri, aprecieri și *retweet*-uri colectate de la fiecare persoană care a atins vreodată un computer conectat în rețea – sunt cernute prin algoritmi brevetați de „învățare automată" pentru a microținti totul, de la reclamele pe care le vedem până la știrile (false sau nu) pe care le întâlnim când navigăm pe internet.

La începutul anilor cincizeci, Wiener a propus ca cercetătorii să studieze structurile și limitările furnicilor – față de ale oamenilor – pentru ca mașinile să obțină într-o bună zi „aproape indefinita extindere intelectuală" pe care oamenii o pot atinge (și insectele nu). A găsit consolare în ideea că mașinile ar putea ajunge să ne domine doar „în ultimele stadii ale entropiei crescânde", când „diferențele statistice dintre indivizi vor fi nule". Algoritmii de minare a datelor de astăzi dau peste cap teoria lui Wiener. Produc profit prin exploatarea creierului nostru reptilian, și nu prin imitarea cortexurilor noastre cerebrale, culegând informații din toate fluxurile noastre de clicuri făcute noaptea târziu, rătăcite pe bloguri și căutătoare de plăceri – valorificând exact micile și rezidualele „diferențe statistice dintre indivizi".

Cu siguranță, câteva realizări recente din domeniul inteligenței artificiale au fost remarcabil de impresionante. Computerele pot produce acum opere de artă vizuală și compoziții muzicale asemănătoare celor ale maeștrilor recunoscuți și creând exact tipul de

„informație" pe care Wiener o valoriza cel mai mult. Dar, de departe, cel mai mare impact asupra societății până astăzi a venit de la colectarea și manipularea informațiilor de tip Shannon, care au dat o nouă formă obiceiurilor noastre de cumpărare, participării politice, relațiilor personale, așteptărilor privind confidențialitatea și multe altele.

În ce ar putea evolua „învățarea profundă" dacă moneda fundamentală devine „informația" așa cum a definit-o Wiener? Cum s-ar schimba domeniul dacă ar fi reanimat de convingerile morale profunde ale lui Wiener, informate fiind de îngrijorările sale vizionare cu privire la militarismul frenetic, setea nestăvilită de profit a corporațiilor, caracteristicile autolimitante ale secretelor și reducerea expresiei umane la bunuri de consum interschimbabile? Poate că „învățarea profundă" ar putea deveni cultivarea informației semnificative în locul urmăririi neîncetate a biților potenți, chiar dacă sunt lipsiți de sens.

Capitolul 16
Scalarea

Neil Gershenfeld

Neil Gershenfeld este fizician și director al Centrului pentru Biți și Atomi de la MIT. Este autorul cărților *FAB*, coautorul (împreună cu Alan Gershenfeld și Joel Cutcher-Gershenfeld) cărții *Designing Reality* și fondatorul rețelei globale Fab Lab.

În sus-menționata discuție din Connecticut despre *The Human Use of Human Beings*, Neil Gerhenfeld a oferit o perspectivă proaspătă, spunând că nu i-a plăcut deloc cartea. Observația sa a fost întâmpinată cu râsete generale – la fel și observația sa că informatica a fost unul dintre cele mai rele lucruri care s-a întâmplat computerelor sau științei. Controversa sa generală a fost că Wiener a ratat implicațiile revoluției digitale care avea loc în jurul său, deși unii ar spune că această acuzație nu poate fi făcută despre o persoană de la cel mai jos nivel și căreia îi lipsește clarviziunea.

„Coada care scutură câinele vieții mele", ne-a spus el, „a fost Fab Labs și mișcarea ei de creare, și, când Wiener vorbește despre amenințarea automatizării, scapă din

vedere reciproca, și anume că accesul la mijloacele de automatizare îi poate împuternici pe oameni, iar, în Fab Labs, colțul în care am fost implicat, acest lucru este exponențial".

În 2003, i-am făcut o vizită lui Neil la MIT, unde conduce Centrul pentru Biți și Atomi. Câteva ore mai târziu, am ieșit din ceea ce fusese o afișare exuberantă de lucruri foarte ciudate. Mi-a arătat activitatea unuia dintre studenții săi la popularul curs pe care-l ținea despre crearea rapidă a prototipurilor (Cum să faci aproape orice), un sculptor fără studii de inginerie care a făcut un spațiu personal portabil de țipat care îți salvează țipetele și le poți asculta mai târziu. Alt student de la cursul său a făcut un browser web care permite papagalilor să navigheze pe internet. Neil însuși făcea cercetări fundamentale pe harta care duce la acel vis science-fiction, „replicatorul universal". A fost o vizită pentru care am avut nevoie de câțiva ani s-o înțeleg.

Neil administrează rețeaua globală Fab Labs – sisteme de producție la scară mică, având la bază tehnologii digitale, ce oferă oamenilor mijloacele necesare pentru a construi ce vor ei. În calitate de guru al acestei mișcări, care unește comunicarea digitală și tehnica de calcul cu procesul de producție, uneori se simte în afara dezbaterii aprinse privind siguranța inteligenței artificiale. „Abilitatea mea de a face cercetări se bazează pe instrumente care-mi augmentează capacitățile", spune el. „A întreba dacă sunt inteligente sau nu este la fel de util ca a întreba cum știu că exist – distractiv din punct de vedere filosofic, dar nu poate fi testat empiric." Ce îl interesează pe el este „cum interacționează biții și atomii – granița dintre digital și fizic. Din punct de vedere științific, este cel mai interesant lucru pe care îl știu".

Discuțiile pe tema inteligenței artificiale au fost dubios de anistorice. Ar putea fi descrise mai bine ca maniaco-depresive; în funcție de cum numărăm, ne aflăm acum în al cincilea ciclu de expansiune și contracție. Aceste pendulări maschează continuitatea din progresul subiacent și implicațiile punctului către care se îndreaptă.

Ciclurile au venit în valuri de aproximativ zece ani. Mai întâi, au fost computerele *mainframe*, care, prin simpla lor existență, urmau să automatizeze munca. Ideea s-a izbit de zidul realității când s-a constat că este greu să scrii programe care să efectueze sarcini ușor de realizat pentru oameni. Apoi au venit experții de sisteme, care urmau să codifice și apoi să-i înlocuiască pe experții în cunoștințe. Aceștia au dat de greu în asamblarea cunoștințelor și raționării cu privire la cazurile care nu erau deja acoperite. Perceptronii căutau să evite aceste probleme modelând modul în care învață creierul, dar nu au făcut o treabă prea bună. Perceptronii cu mai multe niveluri puteau să gestioneze probleme-test care declanșau acele rețele simple, dar demonstrațiile lor nu s-au descurcat cu probleme nestructurate, reale. Acum ne aflăm în era învățării profunde, care a reușit în privința multor promisiuni timpurii ale inteligenței artificiale, dar într-un mod considerat greu de înțeles, cu consecințe care variază de la amenințări intelectuale până la amenințări existențiale.

Fiecare dintre aceste etape era anunțată ca un avans revoluționar asupra limitărilor predecesorilor, însă toate trebuiau să facă

practic același lucru: să facă inferențe din observații. Modul cum relaționează aceste abordări poate fi înțeles din cum pot fi scalate – mai exact, cum depinde performanța lor de dificultatea problemei pe care încearcă să o rezolve. Atât întrerupătorul becului, cât și mașina autonomă trebuie să stabilească intențiile operatorilor, dar primul are doar două opțiuni din care să aleagă, în timp ce mașina are mult mai multe. Etapele boomului inteligenței artificiale au început cu exemple promițătoare în domenii limitate; fazele de extindere au venit cu nereușita demonstrației de a gestiona complexitatea problemelor practice mai puțin structurate.

Mai puțin vizibil este progresul continuu pe care l-am făcut în domeniul scalării. Acest progres are la bază distincția tehnologică dintre funcțiile liniare și cele exponențiale – o distincție care devenea evidentă la începuturile inteligenței artificiale, dar cu implicații pentru inteligența artificială care nu au fost apreciate decât mulți ani mai târziu.

Într-unul dintre documentele fondatoare despre studierea mașinilor inteligente, *The Human Use of Human Beings*, Norbert Wiener face o treabă remarcabilă în identificarea celor mai semnificative tendințe care urmau să ia naștere după ce a scris cartea, observându-i în același timp și pe oamenii responsabili de ele, apoi nereușind niciodată să recunoască de ce munca acestora se va dovedi atât de importantă. Wiener este considerat a fi inventatorul domeniului ciberneticii; nu am înțeles niciodată ce-i asta, dar ceea ce lipsește din carte este tocmai nucleul progresului inteligenței artificiale. Această istorie contează, pentru că ecourile sale au continuat până în prezent.

Claude Shannon apare și el în carte, într-un rol secundar, în contextul ideilor sale despre perspectivele unui computer capabil să joace șah. Shannon făcea ceva mult mai semnificativ decât să speculeze la vremea aceea: turna fundația unei revoluții digitale. Absolvent al Institutului Tehnologic din Massachusetts, a lucrat pentru Vannevar Bush la Analizorul Diferențial. Acesta a fost unul dintre ultimele mari computere analogice, o cameră plină de roți și axuri. Frustrarea lui Shannon în ceea ce privește dificultatea

rezolvării problemelor în acest mod l-a făcut să scrie în 1937 probabil cea mai bună disertație de masterat din toate timpurile. În aceasta, a arătat cum circuitele electrice pot fi proiectate să evalueze expresii logice arbitrare, introducând baza pentru logica digitală universală.

După MIT, Shannon a studiat comunicațiile la Bell Labs. Apelurile telefonice analogice se deteriorau proporțional cu distanța; cu cât ajungeau mai departe, cu atât sunau mai rău. Decât să continuăm să le îmbunătățim incremental, Shannon a arătat în 1948 că, prin comunicare cu simboluri mai degrabă decât cu cantități continue, comportamentul este foarte diferit. Convertirea formelor de undă ale limbajului la valorile binare de 1 și 0 este un exemplu, dar multe alte seturi de simboluri pot fi (și sunt) utilizate în comunicări digitale. Ce contează nu sunt simbolurile particulare, ci mai degrabă capacitatea de a detecta și corecta erori. Shannon a descoperit că, dacă zgomotul este peste o anumită valoare (care depinde de designul sistemului), atunci cu siguranță vor apărea erori. Dar, dacă zgomotul este sub un anumit nivel, atunci o creștere liniară în resursele fizice ce reprezintă simbolul determină o scădere exponențială a probabilității de a face o greșeală în receptarea corectă a simbolului. Această relație a fost prima pentru ceea ce numim în prezent o teoremă a valorii de prag.

O astfel de scalare cade atât de rapid, încât probabilitatea unei erori poate fi atât de mică încât efectiv să nu apară niciodată. Fiecare simbol trimis multiplică, și nu adaugă certitudinea, astfel încât probabilitatea unei greșeli să scadă de la 0,1 la 0,01 la 0,001 și așa mai departe. Această scădere exponențială în erorile de comunicare a făcut posibilă o creștere exponențială a capacității rețelelor de comunicare. Și acest lucru a rezolvat în cele din urmă problema originii cunoștințelor unui sistem de inteligență artificială.

Mulți ani, cel mai rapid mod de a îmbunătăți o tehnică de calcul era să nu faci nimic – doar să aștepți computerele să devină mai rapide. În același mod, au fost ani întregi de proiecte de inteligență artificială care urmăreau să acumuleze cunoștințe de

zi cu zi introducând laborios bucăți de informație. Acest model nu se putea scala; putea progresa doar proporțional cu numărul de oameni care introduceau datele. Dar când apelurile telefonice, articolele de ziar și mesajele de e-mail s-au mutat pe internet, toată lumea care făcea oricare dintre cele menționate a devenit automat un generator de date. Rezultatul a fost o rată de acumulare a cunoștințelor exponențială, și nu liniară.

Și John von Neumann are un rol secundar în *The Human Use of Human Beings* în ceea ce privește teoria jocurilor. Ce a ratat Wiener aici este rolul seminal al lui Neumann în digitalizarea tehnicii de calcul. În timp ce comunicările analogice se degradau proporțional cu distanța, tehnica de calcul analogică (de exemplu, Analizorul Diferențial) se degrada în timp, acumulând erori pe măsură ce progresa. Von Neumann a prezentat în 1952 un rezultat ce corespundea cu al lui Shannon în ceea ce privește tehnica de calcul (s-au întâlnit la Institutul de Studii Avansate din Princeton), arătând că este posibilă o tehnică de calcul fiabilă cu un dispozitiv de calcul nefiabil utilizând simboluri în locul cantităților continue. Acesta era, din nou, un argument pentru scalare cu o creștere liniară a resurselor fizice ce reprezintă simbolul și rezultând într-o reducere exponențială a ratei erorilor, atât timp cât zgomotul era sub un anumit prag. Asta-i ceea ce face posibilă integrarea unui miliard de tranzistori pe un cip de computer, cu ultimul fiind la fel de util ca primul. Această relație a condus la o creștere exponențială în performanța de calcul, ceea ce a rezolvat o a doua problemă a inteligenței artificiale: prelucrarea unui volum de date aflat într-o creștere exponențială.

A treia problemă pe care a rezolvat-o scalarea pentru inteligența artificială a fost structurarea regulilor de raționare fără a fi nevoie de angajarea unui programator pentru fiecare problemă. Wiener a recunoscut rolul *feedback*-ului în învățarea automată, dar a ratat rolul-cheie al reprezentării. Nu este posibil să stochezi toate imaginile posibile într-o mașină autonomă sau toate sunetele posibile într-un computer conversațional; trebuie să poată generaliza din experiență. Partea „profundă" din învățarea profundă nu

se referă la profunzimea (sperată a) observațiilor, ci la profunzimea straturilor de rețele matematice utilizate pentru a face predicții. S-a dovedit că o creștere liniară în complexitatea rețelei conduce la o creștere exponențială în puterea de expresie a rețelei.

Dacă-ți pierzi cheile într-o cameră, le poți căuta. Dacă nu ești sigur în ce cameră se află, trebuie să cauți în toate camerele din clădire. Dacă nu ești sigur în ce clădire se află, trebuie să cauți în toate clădirile din oraș. Dacă nu știi sigur în ce oraș te afli, trebuie să cauți în toate camerele din toate clădirile din toate orașele. În inteligența artificială, găsirea cheilor corespunde unor lucruri precum o mașină care urmează în siguranță un drum sau un computer care interpretează corect o comandă vorbită, iar camerele, clădirile și orașele corespund tuturor opțiunilor care trebuie luate în considerare. Acesta este numit blestemul dimensionalității.

Soluția blestemului dimensionalității a venit din folosirea informațiilor despre problemă pentru a restricționa căutarea. Algoritmii de căutare în sine nu sunt noi, dar, când sunt aplicați la o rețea cu învățare profundă, construiesc adaptiv reprezentări pentru locul în care ar trebui să caute. Prețul acestui lucru este că nu mai este posibil să rezolvi exact pentru cel mai bun răspuns la o problemă, dar, de obicei, este necesar doar un răspuns care să fie suficient de bun.

Luate împreună, nu ar trebui să fie surprinzător că aceste legi de scalare au permis mașinilor să devină efectiv la fel de capabile precum stadiile corespunzătoare ale complexității biologice. Rețelele neuronale au pornit inițial cu obiectivul de a modela cum funcționează creierul. Acel obiectiv a fost abandonat, deoarece au evoluat în abstracții matematice care nu mai aveau nicio legătură cu modul în care funcționează neuronii. Dar acum a apărut un soi de convergență care poate fi considerată o biologie cu inginerie normală, și nu inginerie inversă, deoarece rezultatele învățării profunde oglindesc nivelurile și regiunile creierului.

Unul dintre cele mai dificile proiecte de cercetare pe care le-am gestionat a pus laolaltă ceea ce numim acum cercetători de date și pionieri ai inteligenței artificiale. A fost o experiență mizerabilă

de mutare a barelor porții. Cât timp cercetătorii de date progresau în rezolvarea problemelor vechi ridicate de inteligența artificială, se considera că nu contează, pentru că nu era realizat prin salturi în înțelegerea soluțiilor. Care este valoarea unui computer care joacă șah dacă nu poți să explici cum joacă șah?

Răspunsul, bineînțeles, este că poate juca șah. Există cercetări emergente interesante care aplică inteligența artificială pe inteligența artificială – adică instruiesc rețele ca să explice cum funcționează. Dar atât creierele, cât și cipurile de computer sunt greu de înțeles dacă le privești mecanismele; se interpretează ușor doar prin observarea interfețelor interne. Ajungem să avem încredere (sau nu) în creiere și cipuri de computer în baza unor experiențe și a unor teste, și nu în funcție de explicațiile pe care le primim pentru cum funcționează.

Multe branșe de proiectare fac o tranziție de la ceea ce este numit design imperativ la ceea ce este numit design declarativ sau generativ. Asta înseamnă că, în loc de a proiecta explicit un sistem cu instrumente ca fișierele CAD, scheme de circuite și cod de computer, descrii ce vrei să facă sistemul și apoi este efectuată o căutare automată pentru designurile care satisfac obiectivele și restricțiile tale. Această abordare devine necesară pe măsură ce complexitatea designului depășește ceea ce poate fi înțeles de un proiectant uman. Deși sună ca un risc, înțelegerea omenească vine cu propriile sale limitări; proiectarea în inginerie este presărată cu ceea ce au părut a fi observații bune, dar au avut consecințe negative. Designul declarativ se bazează pe toate progresele din domeniul inteligenței artificiale, plus fidelitatea din ce în ce mai bună a simulărilor care testează virtual designurile.

Mama tuturor problemelor de design este cea care ne-a creat pe noi. Modul în care suntem proiectați rezidă în una dintre cele mai vechi și bine conservate părți ale genomului, numită genele Hox. Acestea sunt gene care reglează genele, în ceea ce se numesc programe de dezvoltare. Nimic din genomul nostru nu memorează structura corpului nostru; genomul memorează, mai degrabă, o serie de etape care urmăresc acel rezultat în corpul nostru.

Aceasta este o paralelă exactă pentru cum se fac căutările în inteligența artificială. Există prea multe planuri posibile pentru corpuri prin care să caute și cele mai multe modificări ar fi fie neimportante, fie fatale. Genele Hox sunt o reprezentare a unui loc productiv pentru căutare evoluționistă. Sunt un fel de inteligență naturală la nivel molecular.

Inteligența artificială are o problemă minte–corp în sensul că nu are corp. Mare parte a muncii inteligenței artificiale este făcută în *cloud*, ce rulează pe mașini virtuale în centre de computere unde sunt direcționate datele. Propria noastră inteligență este rezultatul unui algoritm de căutare (evoluția) care ne-a putut schimba forma fizică, precum și programarea – acestea sunt legate inextricabil. Dacă istoria inteligenței artificiale poate fi înțeleasă ca rezultatul legilor de scalare, mai degrabă decât o succesiune de mode, atunci viitorul său poate fi văzut în același mod. Ce este acum digitalizat, după comunicație și tehnică de calcul, este producția, aducerea capacității de programare a biților în lumea atomilor. Digitalizând nu doar designurile, ci și construcția materialelor, aceleași lecții pe care le-am învățat de la von Neumann și Shannon se aplică și complexității cu creștere exponențială a proceselor de producție.

Am definit materialele digitale ca fiind cele construite dintr-un set discret de părți reversibil combinat cu un set discret de poziții și orientări relative. Aceste atribute permit determinarea geometriei globale din restricții locale, detectarea și corectarea erorilor de asamblare, combinarea materialelor eterogene și demontarea structurilor când nu mai sunt necesare, în loc de distrugerea lor. Aminoacizii ce reprezintă temelia vieții și piesele Lego ce reprezintă fundația jocului împărtășesc aceste proprietăți.

Ce este interesant cu privire la aminoacizi este că nu sunt interesanți. Au atribute care sunt tipice, dar nu neobișnuite, cum ar fi să poată atrage sau respinge apa. Dar nu mai mult de douăzeci de tipuri dintre ei sunt suficiente ca să-ți dea viață. În anumite feluri, douăzeci și ceva de tipuri de piese de material digital – conductoare, izolante, rigide, flexibile, magnetice etc. – sunt

suficiente pentru a asambla un interval de funcții care participă la fabricarea tehnologiilor moderne, precum roboții și computerele.

Legătura dintre tehnică de calcul și fabricație a fost prefigurată chiar de acei pionieri a căror muncă este fundația edificiului tehnicii de calcul. Wiener a făcut aluzie la ea corelând transportul de material cu transferul de mesaje. John von Neumann este creditat ca inventator al arhitecturii computerelor moderne, subiect despre care, de fapt, a scris foarte puțin; ultimul lucru pe care l-a studiat și despre care a scris frumos și pe îndelete au fost sistemele cu reproducere automată. Ca o abstractizare a vieții, a modelat o mașină care poate comunica o tehnică de calcul ce se construiește pe sine. Și ultimul lucru pe care l-a studiat Alan Turing, care este creditat cu cadrul teoretic al informaticii, a fost cum pot instrucțiunile din gene să dea naștere formelor fizice. Aceste întrebări vizează o temă ce lipsește din educația informatică obișnuită: configurația fizică a unei tehnici de calcul.

Von Neumann și Turing și-au pus întrebările ca studii teoretice, pentru că erau dincolo de posibilitățile tehnologiei din vremea lor și nu puteau fi puse în practică. Dar, odată cu convergența comunicațiilor și tehnicii de calcul cu producția, aceste investigații devin acum accesibile experimental. În laboratorul meu ne concentrăm pe asamblarea unui robot de asamblare care se poate asambla pe sine din părțile componente, dar colaborăm și cu alte companii pentru a dezvolta celule sintetice.

Ideea automatelor care se pot reproduce automat și fizic are un potențial mult mai înspăimântător decât temerile legate de inteligența artificială scăpată de sub control, pentru că mută inteligența în afară, unde trăim noi. Poate fi o traiectorie care ne va duce la stăpânii robotici Skynet din filmul *Terminator*. Dar există și idei mai luminoase, pentru că abilitatea de a programa atomi la fel de bine cum programăm biții permite designurilor să fie împărtășite global, producând în același timp local energie, alimente și adăpost – toate acestea sunt emergente sub formă de aplicații noi și interesante al fabricației digitale. Wiener își făcea probleme în legătură cu viitorul muncii, dar nu a pus la îndoială

presupunerile implicite cu privire la natura muncii care sunt provocate atunci când consumul poate fi înlocuit prin creație.

Istoria ne învață că nu prevalează nici scenariile utopice, nici cele distopice; în general, orbecăim undeva între. Dar istoria ne învață și că nu trebuie să așteptăm istoria. Gordon Moore, în 1965, a putut utiliza cinci ani de dublare a specificațiilor circuitelor integrate pentru a proiecta ceea ce s-au dovedit a fi cincizeci de ani de îmbunătățiri exponențiale în tehnologiile digitale. Am petrecut mulți din acei ani răspunzând la, în loc să anticipăm, implicațiile sale. Avem mai multe date disponibile acum decât avea Gordon Moore pentru a proiecta cincizeci de ani de dublare a performanțelor producției digitale. Înarmați cu beneficiile retrospectivei, ar trebui să fie posibil să evităm excesele tehnicii de calcul digitale și comunicațiilor de data aceasta și, pentru început, să ne ocupăm de problemele de acces și alfabetizare.

Dacă mișcarea celor ce produc este vestitorul unei a treia revoluții digitale, succesul unei inteligențe artificiale în îndeplinirea multor obiective inițiale poate fi văzut ca realizarea principală a primelor două revoluții digitale. Deși producția de mașini și modul în care gândesc mașinile par să fie două tendințe separate, în viitor depind una de cealaltă. Aceleași tendințe de scalare care au făcut posibilă inteligența artificială sugerează că mania curentă este o fază trecătoare, care va fi urmată de ceva și mai important: îmbinarea inteligenței artificiale cu inteligența naturală.

A fost un avans când atomii au format molecule, moleculele au format organite celulare, organitele au format celule, celulele au format organe, organele au format organisme, organismele au format familii, familiile au format societăți și societățile au format civilizații. Acest ciclu evoluționist la scară mare poate fi închis acum cu atomi care aranjează biți aranjatori de atomi.

Capitolul 17
Primele inteligențe automate

W. Daniel Hills

W. Daniel „Danny" Hillis este inventator, antreprenor și informatician, profesor „Judecător Widney" de inginerie și medicină la USC și autorul cărții *The Pattern on the Stone: The Simple Ideas That Make Computers Work.*

Cât timp Danny Hillis a fost student la MIT, a construit un computer din jucării de lemn. Are aproximativ 10 000 de piese de lemn, joacă X și 0 și nu pierde niciodată; acum se află la Muzeul de Istorie a Computerelor din Mountain View, California.

După ce a absolvit și lucra la Laboratorul de inteligență artificială și informatică de la MIT, la începutul anilor 1980, Danny a conceput un computer masiv care funcționa în paralel, cu 64 000 de procesoare. L-a numit Mașina Conexiunilor și a înființat ceea ce a fost probabil prima companie de inteligență artificială – Thinking Machines Corporation – pentru a-l produce și a-l comercializa. Aceasta, în ciuda unui prânz pe care l-a luat împreună cu Richard Feynman în timpul căruia renumitul fizician a remarcat, „asta e de-a dreptul cea mai nătângă idee pe care

am auzit-o vreodată". Poate „în ciuda" este cuvântul nepotrivit, de vreme ce Feynman avea o predilecție bine-cunoscută de a se juca cu ideile nătânge. La eveniment, s-a prezentat în ziua înființării companiei și a rămas în comitetul de conducere, pentru activități pe timpul verii și sarcini speciale, aducând astfel contribuții importante la activitatea sa.

Danny a înființat de atunci mai multe companii de tehnologie, dintre care cea mai nouă este Applied Invention, care face parteneriate cu întreprinderi comerciale pentru a dezvolta soluții tehnologice pentru cele mai dificile probleme pe care le au. Deține sute de brevete în SUA, care acoperă computerele în paralel, interfețele tactile, matricele de discuri, metode de împiedicare a falsurilor și multe dispozitive mecanice și electrice. Imaginația sa este, aparent, fără limite și schițează aici câteva scenarii posibile ce vor lua naștere din urmărirea unei inteligențe artificiale din ce în ce mai bune.

„Mașinile noastre gânditoare sunt mai mult decât niște simple metafore", spune el. „Întrebarea nu este dacă vor fi suficient de puternice pentru a ne face rău (pentru că vor fi) sau dacă vor acționa întotdeauna în interesele noastre (pentru că nu o vor face), ci dacă, pe termen lung, ne vor ajuta să ne găsim calea – unde vom ajunge pe continuumul Panaceu/Apocalipsă."

Am vorbit deapre mașini, dar nu doar despre mașini cu creiere de alamă și mușchi de fier. Când atomii umani sunt croșetați într-un organism în care sunt folosiți, nu în dreptul lor deplin de ființe umane responsabile, ci ca rotițe, manete și tije, contează foarte puțin că materia lor primă sunt carnea și oasele. Ce este utilizat ca element într-o mașină este, de fapt, un element într-o mașină. *Indiferent dacă încredințăm deciziile noastre mașinilor de metal sau mașinilor din carne și oase care sunt birourile, laboratoarele vaste, armatele și corporațiile, nu vom primi niciodată răspunsul corect la întrebările noastre decât dacă punem întrebările potrivite... Este foarte târziu și alegerea între bine și rău bate la ușă.*

— Norbert Wiener, *The Human Use of Human Beings*

Norbert Wiener a fost un profet în ceea ce privește recunoașterea potențialelor pericole create de mașinile inteligente emergente. Eu consider că a fost și mai profetic recunoscând că primele inteligențe artificiale erau deja emergente. A identificat corect corporațiile și birourile pe care le numea „mașini în carne și oase" ca fiind primele mașini inteligente. A anticipat pericolele creării superinteligențelor artificiale care ar putea avea obiective ce nu sunt neapărat aliniate cu ale noastre.

Ce este clar acum, indiferent dacă era sau nu clar și pentru Wiener, este că aceste superinteligențe organizaționale nu sunt făcute doar din oameni, ci sunt hibrizi între oameni și tehnologiile informatice care le permit coordonarea. Chiar și pe vremea lui Wiener, „birourile, laboratoarele vaste, armatele și corporațiile" nu puteau funcționa fără telefoane, telegrafe, radiouri și tabulatoare. Astăzi nu ar putea funcționa fără rețele de computere, baze de date și sisteme de asistență în luarea deciziilor. Aceste inteligențe hibride sunt rețele de oameni augmentate tehnologic. Aceste inteligențe artificiale au puteri supraomenești. Pot ști mai mult decât indivizii în parte, pot simți mai mult, pot face analize mai complicate și planuri mai complexe. Pot avea mult mai multe resurse și mai multă putere decât orice individ în parte.

Deși nu percepem întotdeauna acest lucru, superinteligențele hibride cum ar fi statele naționale și corporațiile au propriile lor obiective emergente. Deși sunt construite de și pentru oameni, adesea acționează ca niște entități inteligente independente și acțiunile lor nu sunt aliniate întotdeauna cu interesele oamenilor care le-au creat. Statul național nu este întotdeauna pentru cetățean și nici compania pentru acționar. La fel, organizațiile nonprofit, ordinele religioase și partidele politice funcționează întotdeauna pentru promovarea principiilor fondatoare. Intuitiv, recunoaștem că acțiunile lor sunt ghidate de obiective interne, motiv pentru care le personificăm, atât legal, cât și în obiceiurile noastre de gândire. Când vorbim despre „ce vrea China" sau „ce încearcă General Motors să facă", nu vorbim în metafore. Aceste organisme acționează ca niște inteligențe care percep, iau decizii și acționează. La fel ca obiectivele oamenilor, obiectivele organizațiilor sunt complexe și adesea contradictorii, dar sunt obiective adevărate în sensul că direcționează acțiunea. Aceste obiective depind oarecum de obiectivele oamenilor din organizație, dar nu sunt identice.

Orice american știe cât de slabă este legătura dintre acțiunile guvernului american și aspirațiile diverse și adesea contradictorii ale cetățenilor săi. Același lucru este adevărat și pentru corporații.

Corporațiile orientate către profit servesc nominal mai multor constituenți, inclusiv acționarilor, directorilor de departamente, angajaților și clienților. Aceste corporații diferă în modul în care jonglează cu loialitatea și se comportă adesea în moduri care nu servesc niciunuia dintre constituenții lor. „Neuronii" care transportă gândirea corporatistă nu sunt doar angajații umani sau tehnologiile care îi conectează; sunt, de asemenea, codați în regulile, structurile de stimulente, cultura și obiceiurile procedurale ale corporației. Obiectivele corporatiste emergente nu reflectă întotdeauna valorile oamenilor care le implementează. De exemplu, o companie petrolieră condusă de și în care lucrează oameni cărora le pasă de mediu poate avea structuri de stimulente sau reguli care o vor face să compromită siguranța mediului de dragul câștigurilor corporative. Intențiile bune ale componenților nu sunt o garanție pentru comportamentul bun emergent al sistemului.

Guvernele și corporațiile, ambele compuse parțial din oameni, sunt motivate natural să pară cel puțin că împărtășesc obiectivele oamenilor de care depind. Nu ar putea funcționa fără oameni, deci trebuie să-i țină cooperanți. Când astfel de organizații par să se comporte altruist, aceasta este adesea o parte importantă din motivație. Am complimentat la un moment dat CEO-ul unei corporații mari pentru contribuția pe care a făcut-o compania sa pentru un efort de întrajutorare umanitară. CEO-ul a răspuns fără urmă de ironie: „Da. Am hotărât să facem mai multe lucruri ca acela pentru a face marca noastră mai plăcută". Indivizii care compun o superinteligență hibridă pot exercita uneori o influență de „umanizare" – de exemplu, un angajat poate încălca regulile companiei pentru a răspunde nevoilor altui om. Angajatul poate acționa din empatie veritabilă, dar nu trebuie să atribuim o astfel de empatie superinteligenței în sine. Mașinile hibride au obiective și cetățenii/clienții/angajații lor sunt câteva dintre resursele pe care le folosesc pentru a le atinge.

Suntem aproape de a putea construi superinteligențe dintr-o tehnologie a informației pură, fără componente umane. La asta se referă oamenii în general când spun „inteligență artificială". Este

rezonabil să întrebăm care vor fi atitudinile acestor mașini superinteligente ipotetice față de oameni. Îi vor vedea și ele pe oameni ca pe niște resurse utile și importanța păstrării unor relații bune? Vor fi construite astfel încât să aibă obiective aliniate cu ale noastre? O superinteligență ar sesiza importanța acestor întrebări? Care sunt „întrebările potrivite" pe care ar trebui să ni le punem? Cred că una dintre cele mai importante este următoarea: ce relație vor avea diversele superinteligențe una cu cealaltă?

Este interesant să ne gândim cum se descurcă în prezent superinteligențele hibride cu conflictele dintre ele. Astăzi o mare parte dintre puterile depline stă în brațele statelor naționale, care pretind autoritate peste un petic de pământ. Indiferent dacă sunt optimizate pentru a acționa în interesul cetățenilor lor sau în cel al unui conducător despotic, statele naționale își arogă prioritate asupra dorințelor și obiectivelor altor inteligențe din zona geografică controlată. Pretind monopol asupra folosirii forței și recunosc doar alte state naționale ca egali. Sunt dispuse, dacă este cazul, să solicite sacrificii mari din partea cetățenilor pentru a-și asigura autoritatea, chiar până la sacrificarea vieților cetățenilor.

Această împărțire geografică a autorității avea sens când majoritatea actorilor erau oameni care-și petreceau întreaga viață într-un singur stat național, dar acum, când actorii importanți includ inteligențe hibride distribuite geografic, cum ar fi corporațiile multinaționale, această logică este mai puțin evidentă. Astăzi trăim într-o perioadă de tranziție complexă, când superinteligențele distribuite încă se bazează în mare parte pe statele naționale pentru a rezolva problemele care iau naștere între ele. Adesea aceste probleme sunt rezolvate diferit în diferite jurisdicții. Devine din ce în ce mai greu să atribuim persoanele individuale statelor naționale: călătorii internaționali care trăiesc și lucrează în afara țării lor de origine, refugiații și imigranții (cu sau fără acte) încă sunt priviți ca niște excepții stingheritoare. Superinteligențele construite strict din tehnologia informației se vor dovedi și mai stingheritoare pentru sistemul teritorial de autoritate, pentru că nu va mai exista niciun motiv să fie legate de

resursele fizice dintr-o singură țară – de altfel, de nicio resursă fizică anume. O inteligență artificială poate exista bine-mersi „în *cloud*", și nu într-un anumit loc fizic.

Pot să-mi imaginez cel puțin patru scenarii pentru modul în care superinteligențele-mașină vor relaționa cu superinteligențele hibride.

Într-un scenariu evident, mai multe inteligențe mașină vor fi controlate și aliate în cele din urmă cu statele naționale individuale. În acest scenariu stat/inteligență artificială, ne putem imagina superinteligențele artificiale americane și chineze luptându-se pentru resurse în numele statului lor. Într-un anumit sens, aceste inteligențe artificiale ar fi cetățeni ai statelor lor naționale în modul în care multe corporații comerciale acționează adesea drept „cetățeni corporativi" astăzi. În acest scenariu, statele naționale gazdă probabil ar conferi superinteligențelor-mașină resursele de care au nevoie pentru a lucra în avantajul statului. Or, în măsura în care superinteligențele pot influența guvernele lor naționale, vor face asta pentru a-și spori propriile puteri, de exemplu colectând o mare parte din resursele statului. Inteligențele artificiale ale statelor naționale probabil nu vor vrea ca în jurisdicția lor să crească inteligențe artificiale concurente. În acest scenariu, superinteligențele devin o extensie a statului și viceversa.

Scenariul stat/inteligență artificială pare plauzibil, dar nu este cursul nostru curent. Cele mai puternice inteligențe artificiale ale noastre și cele care se îmbunătățesc cel mai rapid sunt cele controlate de corporațiile orientate pe profit. Acesta este scenariul corporație/inteligență artificială, în care echilibrul puterii între statele naționale și corporații se inversează. Astăzi cele mai puternice și inteligente colecții de mașini sunt probabil deținute de Google, dar companii precum Amazon, Baidu, Microsoft, Facebook, Apple și IBM nu sunt, probabil, cu mult în urmă. Aceste companii văd toate un imperativ comercial în a construi inteligențe artificiale ele însele. Este ușor să ne imaginăm un viitor în care corporațiile își construiesc propriile inteligențe-mașină, protejate de *firewall*-uri și împiedicând mașinile să profite una de

cunoștințele celeilalte. Aceste mașini vor fi proiectate astfel încât să aibă obiective aliniate cu cele ale corporațiilor. Dacă această aliniere este eficientă, statele naționale ar putea continua să rămână în urmă în dezvoltarea propriilor capacități de inteligență artificială depinzând în schimb de „cetățenii lor corporatiști" pentru a face acest lucru în locul lor. În măsura în care corporațiile își controlează cu succes obiectivele, vor deveni mai puternice și mai autonome decât statele naționale.

Alt scenariu, poate cel de care oamenii se tem cel mai mult, este că inteligențele artificiale nu vor fi aliniate cu inteligențele omenești sau cu superinteligențele hibride, ci vor acționa strict în propriul interes. S-ar putea chiar uni într-o superinteligență-mașină unică, de vreme ce nu va exista nicio cerință tehnică pentru ca inteligențele-mașină să-și păstreze identitățile individuale. Atitudinea unei superinteligențe artificiale interesate doar de sine față de superinteligențele hibride va fi probabil una competitivă. Oamenii ar putea fi priviți ca niște obstacole mărunte, ca furnicile la un picnic, dar superinteligențele hibride – cum ar fi corporațiile, religiile organizate și statele naționale – ar putea fi amenințări existențiale. La fel ca superinteligențele hibride, inteligențele artificiale i-ar putea vedea pe oameni în principiu ca pe niște instrumente utile pentru atingerea obiectivelor lor, ca pe niște pioni în competiția lor cu alte superinteligențe. Sau am putea fi pur și simplu irelevanți. Nu este imposibil ca o inteligență mașină să fi luat deja naștere, iar noi să nu o recunoaștem. Poate că nu-și dorește să fie observată sau poate fi atât de străină pentru noi încât să nu o putem percepe. Acest lucru face ca scenariul inteligenței artificiale interesate de sine să fie cel mai greu de imaginat. Cred că versiunile ușor de imaginat, cum ar fi roboții inteligenți umanoizi din science-fiction, sunt cel mai puțin probabile. Cele mai complexe mașini ale noastre, de exemplu internetul, au crescut deja dincolo de înțelegerea detaliată a unui singur om, iar comportamentele lor emergente ar putea fi mult în afara sferei noastre de cunoștințe.

Ultimul scenariu este cel în care inteligențele-mașină nu vor fi aliate una cu cealaltă, ci vor lucra să atingă obiectivele umanității

ca întreg. În acest scenariu optimist, inteligența artificială ne-ar putea ajuta să restaurăm echilibrul puterii între indivizi și corporații, între cetățean și stat. Ne-ar putea ajuta să rezolvăm problemele care au fost create de superinteligențele hibride care subminează obiectivele oamenilor. În acest scenariu, inteligențele artificiale ne vor conferi putere, oferindu-ne acces la capacitatea de procesare și la cunoștințele disponibile în prezent doar statelor și corporațiilor. Prin urmare, ar putea deveni extinderi ale propriilor noastre inteligențe, pentru atingerea obiectivelor noastre umane. Ar putea face inteligențele noastre individuale slabe puternice. Această perspectivă este revigorantă și plauzibilă. Este plauzibilă pentru că avem de ales în ceea ce construim și avem obiceiul de a folosi tehnologia pentru a extinde și augmenta capacitățile umane. La fel cum avioanele ne-au dat aripi și motoarele ne-au dat mușchi să mutăm munții din loc, la fel rețelele noastre de computere ne pot amplifica și extinde mințile. Poate că nu ne înțelegem sau controlăm pe deplin destinul, dar avem șansa să îl orientăm în direcția valorilor noastre. Viitorul nu este ceva ce ni se va întâmpla; este ceva ce vom construi noi.

De ce a văzut Wiener ce nu au văzut alții

> În ingineria electrică, există o divizare care este cunoscută în Germania drept divizarea între tehnica curenților puternici și tehnica curenților slabi și pe care noi o cunoaștem ca distincția dintre ingineria electrică și ingineria telecomunicațiilor. Această divizare separă epoca ce tocmai s-a încheiat și epoca în care trăim acum.
>
> — Norbert Wiener, *Cibernetica sau controlul și comunicațiile la animale și mașini*

Cibernetica este studierea modului în care ceva slab controlează ceva puternic. Gândește-te la metafora definitorie a domeniului: timonierul care ghidează nava cu ajutorul timonei. Obiectivul timonierului este să controleze direcția navei, să o țină

pe cursul potrivit. Informația, mesajul care este transmis timonierului, vine de la busolă sau de la stele, iar timonierul închide ciclul de *feedback* trimițând mesajele de ghidare prin forța fină a mâinii sale către timonă. În această imagine, vedem nava bătută de vânturi și valuri puternice în lumea reală, controlată de sistemul de comunicare al mesajelor din lumea informației.

Însă distincția dintre „real" și „informație" este în cea mai mare parte o diferență în perspectivă. Semnalele care transportă mesaje, cum ar fi lumina stelelor și presiunea mâinii pe timonă, există într-o lume de energie și forțe, la fel și timonierul. Forțele slabe care controlează cârma sunt la fel de reale și de fizice precum forțele puternice ce bat în navă. Dacă schimbăm perspectiva noastră cibernetică de la navă la timonier, presiunea cârmei devine o forță puternică a mușchilor controlată de semnalele slabe ale minții timonierului. Mesajele din mintea timonierului sunt amplificate într-o forță fizică suficient de puternică încât să ghideze nava. Sau, în schimb, putem mări distanța focală a imaginii și putem vedea perspectiva de ansamblu a ciberneticii. Putem vedea nava ca parte dintr-o rețea comercială vastă, parte dintr-un ciclu cu *feedback* ce reglementează prețul mărfurilor prin fluxul de bunuri. În această perspectivă, mica navă este doar un mesager. Astfel că distincția dintre lumea fizică și lumea informațiilor este modul în care descriem relația dintre slab și puternic.

Wiener a ales să vadă lumea din punctul de vedere și la scara omului individual. Ca cibernetician, a preluat perspectiva protagonistului slab integrat într-un sistem puternic, încercând să profite la maximum de puterile sale limitate. A încorporat această perspectivă în propria sa definiție a informației. „Informația", spunea el, „este numele conținutului a ceea ce este schimbat cu lumea exterioară pe măsură ce ne potrivim ei și facem această potrivire resimțită și asupra ei". În cuvintele sale, informația este ceea ce folosim pentru a „trăi efectiv în acel mediu"*. Pentru

* *The Human Use of Human Beings* (Boston: Houghton Mifflin, 1954), 17–18 (n.a.).

Wiener, informația este modalitatea în care slabul face față puternicului. Acest punct de vedere se reflectă și în definiția lui Gregory Bateson a informației ca „diferență ce face diferența", prin care se referea la micile diferențe care determină o diferență mare.

Obiectivul ciberneticii era să creeze un model mic al sistemului folosind „curenții slabi" pentru a amplifica și controla „curenții puternici" din lumea reală. Observația centrală era că o problemă de control putea fi soluționată construind un sistem analog în spațiul de informație al mesajelor și apoi amplificând soluțiile în lumea mai largă a realității. Inerent mișcării unui sistem de control este conceptul de amplificare ce face din mic mare și din slab puternic. Amplificarea permite diferența care face diferența în a face o diferență.

În acest mod de a privi lumea, un sistem de control trebuie să fie la fel de complex ca sistemul pe care îl controlează. Cibernetician W. Ross Ashby a dovedit că acest lucru este adevărat într-un sens matematic și precis, în ceea ce se numește acum Legea lui Ashby privind diversitatea necesară sau, uneori, prima lege a ciberneticii. Legea ne spune că, pentru a controla complet un sistem, controlorul trebuie să fie la fel de complex precum controlatul. Ciberneticienii au tendința să vadă sistemele de control ca un fel de analog pentru sistemele pe care le controlează, ca homunculusul – mica persoană ipotetică din creier care controlează persoana fizică.

Această noțiune de structură analogică este uneori confundată cu noțiunea codării analogice a mesajelor, dar cele două sunt diferite din punct de vedere logic. Norbert Wiener a fost foarte impresionat de analizorul diferențial digital al lui Vannevar Bush, care putea fi reconfigurat pentru a se potrivi structurii oricărei probleme îi era dată spre rezolvare, dar folosea o codare digitală a semnalului. Semnalele puteau fi simplificate pentru a reprezenta deschis distincțiile relevante, permițându-le să fie comunicate și stocate cu o precizie mai mare. În semnalele digitale, trebuia păstrată doar diferența semnalelor care constituiau diferența. Această diferență și codare de semnal este utilizată în prezent

pentru a diferenția între „analog" și „digital". Codarea digitală a semnalului era complet compatibilă cu gândirea cibernetică – de fapt, aceasta o făcea posibilă. Ce era restrictiv pentru cibernetică era presupunerea unei analogii de structură între controlor și controlat. Până în 1930, Kurt Gödel, Alonzo Church și Alan Turing descriseseră cu toții sisteme universale de calcul, în care tehnica de calcul nu necesita nicio analogie structurală pentru funcțiile calculate. Aceste computere universale puteau, de asemenea, să calculeze funcțiile de control.

Analogia de structură dintre controlor și controlat a fost centrală pentru perspectiva ciberneticii. La fel cum codarea digitală comprimă spațiul mesajelor posibile într-o versiune simplificată ce reprezintă doar diferența ce face o diferență, la fel și sistemul de control comprimă spațiul stării sistemului controlat într-un model simplificat ce reflectă doar obiectivele controlorului. Legea lui Ashby nu implică faptul că orice controlor trebuie să modeleze fiecare stare a sistemului, ci doar acele stări care contează pentru avansarea obiectivelor controlorului. Prin urmare, în cibernetică, obiectivul controlorului devine perspectiva din care este văzută lumea.

Norbert Wiener a adoptat perspectiva omului individual relaționând cu organizații vaste și încercând să „trăiască efectiv în acel mediu". A preluat perspectiva celui slab încercând să-l influențeze pe cel puternic. Poate că acesta este motivul pentru care a putut observa obiectivele emergente ale „mașinilor în carne și oase" și anticipa câteva dintre provocările umane ridicate de aceste inteligențe noi – inteligențe mașinice hibride cu obiective proprii.

Capitolul 18
Vor deveni computerele stăpânii noștri

Venki Ramakrishnan

Venki Ramakrishnan este cercetător la Laboratorul de Biologie Moleculară al Consiliului de Cercetare Medicală de la Universitatea Cambridge, laureat al Premiului Nobel în chimie (2009), președintele actual al Societății Regale și autorul cărții *Gene Machine: The Race to Discover the Secrets of the Ribosome.*

Venki Ramakrishnan este un biolog laureat al Premiului Nobel ale cărui contribuții științifice includ activitatea sa în domeniul structurii atomice a ribozomilor – care este, efectiv, o uriașă mașină moleculară care ne citește genele și face proteine. Activitatea sa ar fi fost imposibilă fără computere puternice. Internetul i-a ușurat mult munca și, observă el, a acționat ca un element omogenizator internațional: „Când am crescut în India, dacă voiai să faci rost de o carte, apărea cu șase luni sau chiar un an după ce vedea lumina tiparului în Occident... Revistele ajungeau prin poșta obișnuită după câteva luni. Nu a trebuit să fac față acestei probleme, pentru că am plecat din India la vârsta de nouăsprezece ani, dar cunosc cercetători indieni care au

trebuit să-i facă față. Astăzi, au acces la informație printr-un simplu clic. Mai important, au acces la prelegeri. Îl pot asculta pe Richard Feynman. Acesta ar fi fost un vis al meu din copilărie. Îl pot urmări pur și simplu pe Richard Feynman pe internet. Aceasta este o nivelare foarte puternică în domeniu". Cu toate acestea... „alături de beneficiile internetului, acum a apărut un volum uriaș de zgomot. Apar tot soiul de oameni care perorează în jargon pseudo-științific, promovându-și ideile ca și cum ar fi științifice".

În calitate de președinte al Societății Regale, Venki se îngrijorează și el despre problema mai largă a încrederii: încrederea publică în studiile științifice bazate pe dovezi, dar și încrederea în rândurile oamenilor de știință, consolidată de verificările riguroase ale concluziilor unui coleg de breaslă – încredere care este în pericol de erodare din cauza caracterului de „cutie neagră" al computerelor cu învățare profundă. „Această eroziune va apărea din ce în ce mai mult, pe măsură ce seturile de date se măresc, pe măsură ce vom avea studii generale despre genom, studii pe populație și tot soiul de alte studii", spune el. „Cum vom face față noi, comunitatea științifică, acestui lucru și cum vom comunica publicului o idee despre esența științei, ce este fiabil în știință, ce este nesigur în știință și ce este pur și simplu greșit în știință?"

Un fost coleg al meu, Gérard Bricogne, obișnuia să glumească spunând că inteligența bazată pe carbon este doar un catalizator pentru evoluția inteligenței bazate pe siliciu. Mult timp, atât filmele de la Hollywood, cât și profeții științifici au preconizat capitularea noastră în fața stăpânilor noștri, computere. Cu toții așteptăm singularitatea, care pare întotdeauna să fie după colț.

Într-un anumit sens, computerele au preluat deja controlul, facilitând practic fiecare aspect al vieților noastre – de la banking, călătorii și utilități până la cele mai intime comunicări personale. Pot să-l văd și să vorbesc *gratuit* cu nepotul meu din New York. Îmi amintesc când am văzut prima dată filmul din 1968 intitulat *2001: Odiseea spațială*, publicul a râs când a auzit prețul ridicol de ieftin de 1,7 dolari pentru un apel telefonic cu imagini venite din spațiu, în acel moment un apel la distanță costând 3 dolari pe minut în Statele Unite.

Totuși, comoditatea și puterea computerelor este și ceva asemănător unui târg faustic, deoarece vine la pachet cu pierderea controlului. Computerele ne împiedică să facem lucrurile pe care vrem să le facem. Încearcă să te urci într-un avion dacă ajungi la aeroport și sistemele de computere nu funcționează, așa cum i s-a întâmplat cu puțin timp în urmă companiei British Airways la aeroportul Heathrow. Avioanele, piloții și pasagerii erau toți acolo, chiar și controlul traficului aerian funcționa. Dar niciun avion al companiei aeriene nu putea decola. Computerele ne fac și să facem lucruri pe care *nu vrem* să le facem – generând liste de

e-mailuri și tipărind etichete pentru a ne trimite milioane de e-mailuri nedorite pe care noi, oamenii, trebuie să le sortăm, expediem și ștergem.

Dar nu ați văzut nimic încă. În trecut, am programat computerele folosind algoritmi pe care i-am înțeles cel puțin în principiu. Și, când mașinile au obținut rezultate uimitoare, cum ar fi să-l învingă pe campionul mondial la șah, Garry Kasparov, am putut spune că programele victorioase au fost proiectate cu algoritmi bazați pe propria noastră înțelegere – folosind, în acest caz, experiența și sfaturile celor mai buni jucători de șah. Mașinile erau pur și simplu mai rapide la efectuarea calculelor brute, aveau volume prodigioase de memorie și nu erau pasibile de eroare. Un articol a descris victoria lui Deep Blue nu ca pe cea a unui computer, care era doar o mașină proastă, ci ca victoria a sute de programatori asupra lui Kasparov, un singur om.

Acest mod de programare se modifică dramatic. După o discontinuitate mare, puterea învățării automate a prins aripi. O mare parte din schimbare a venit atunci când programatorii, în loc să încerce să anticipeze și să codeze toate contingențele posibile, au permis computerelor să se instruiască singure în ce privește datele, folosind rețele neuronale profunde bazate pe modele pentru cum învață creierele noastre. Folosesc metode probabilistice pentru „a învăța" din volume mari de date. Computerele pot recunoaște tipare și pot trage concluzii fără ajutorul oamenilor. O metodă foarte puternică pare să fie învățarea cu întărire, prin care computerul învață, fără intrări anterioare, ce variabile sunt importante și câtă pondere să le acorde pentru a atinge un anumit obiectiv. Această metodă imită într-o oarecare măsură modul în care învățăm când suntem copii. Rezultatele acestor abordări noi sunt uimitoare.

Un astfel de program de învățare profundă a fost folosit pentru a învăța un computer să joace Go, un joc despre care doar până acum câțiva ani se credea că nu poate fi jucat de o inteligență artificială pentru că era atât de greu să estimezi cât de bine stai. Se pare că cei mai buni jucători de Go se bazau foarte mult pe

intuiție și pe estimarea poziției, deci se considera că priceperea necesita un anumit tip de inteligență umană. Dar programul AlphaGo produs de DeepMind, după ce a fost instruit din mii de jocuri Go de nivel superior jucate de oameni și apoi de milioane de jocuri pe care le-a jucat cu el însuși, a putut foarte repede să-i bată pe cei mai buni jucători din lume. Și mai uimitor este faptul că programul corespunzător AlphaGo Zero, care a învățat totul de la zero jucând singur, a fost mai puternic decât versiunea instruită inițial pe jocurile cu oameni! Este ca și cum oamenii au împiedicat computerul să ajungă la potențialul său maxim. Aceeași metodă a fost recent generalizată: începând de la zero, în numai douăzeci și patru de ore, un program echivalent de șah de la AlphaZero a reușit să învingă cele mai bune programe de șah „convenționale" de astăzi, care, la rândul lor, au învins cei mai buni oameni.

Progresul nu a fost restricționat la jocuri. Computerele sunt semnificativ mai bune la recunoașterea imaginilor și la sintetizarea vocii decât erau înainte. Pot detecta tumorile din radiografii mai devreme decât majoritatea oamenilor. Diagnosticele medicale și medicamentele personalizate se vor îmbunătăți substanțial. Transportul cu mașinile autonome ne va menține pe toți în siguranță, în medie. Probabil nepotul meu nu va trebui să-și ia niciodată permisul de conducere, pentru că a conduce o mașină va fi cum este astăzi călăritul – un hobby pentru câțiva oameni. Activitățile periculoase cum ar fi minatul și activitățile repetitive și plictisitoare vor fi făcute de computere. Statul va oferi servicii publice mai precise, mai personalizate și mai eficiente. Inteligența artificială poate revoluționa educația analizând toate nevoile individuale ale elevilor și permițând predarea personalizată, astfel încât toți elevii să poată avansa la o rată optimă.

Alături de aceste beneficii uriașe, bineînțeles, vom avea și riscuri alarmante. Cu volumul vast de date personale, computerele vor putea afla mai multe despre noi decât știm noi înșine despre noi; cine deține datele despre noi va deveni o întrebare foarte importantă. În plus, deciziile bazate pe date vor reflecta fără

îndoială predilecțiile sociale: chiar și un așa-zis sistem inteligent neutru conceput pentru prezicerea riscurilor de credit, să spunem, poate trage concluzia că simpla calitate de membru al unui grup minoritar te predispune la incapacitatea de rambursare a împrumutului. Deși acesta este un exemplu evident pe care-l putem corecta, pericolul real este că nu suntem întotdeauna conștienți de predilecțiile din date și le putem perpetua pur și simplu.

De asemenea, învățarea automată poate perpetua propriile noastre prejudecăți. Când Netflix sau Amazon încearcă să-ți spună ce ți-ar plăcea să vizionezi sau să cumperi, aceasta este o aplicație a învățării automate. În prezent, astfel de sugestii sunt uneori ridicole, dar cu timpul și cu mai multe date, vor fi din ce în ce mai precise, fortificându-ne prejudecățile, ce ne place și ce nu ne place. Vom rata oare ocazia întâlnirii aleatorii care ne-ar putea convinge să ne schimbăm părerea prin expunere la idei noi și contradictorii? *Social-media*, dată fiind influența sa asupra alegerilor, este un exemplu destul de clar pentru modul în care poate fi accentuată falia dintre oamenii care se află în părți diferite ale spectrului politic.

Probabil am atins deja etapa în care majoritatea guvernelor nu mai au puterea să reziste la influența combinată a câtorva companii multinaționale puternice care ne controlează pe noi și viitorul nostru digital. Lupta dintre companiile dominante de astăzi este, de fapt, o luptă pentru controlarea datelor. Acestea își vor folosi influența lor enormă pentru a împiedica reglementarea datelor, pentru că sunt interesate de controlul absolut asupra lor. În plus, posedă resursele financiare pentru a-i angaja pe cei mai talentați oameni din domeniu, sporindu-și astfel și mai mult puterea. Ne-am cedat datele personale în schimbul câtorva servicii gratuite precum Gmail și Facebook, dar, după cum evidențiia autorul și jurnalistul John Lanchester în *London Review of Books*, dacă este gratuit, atunci tu ești produsul. Clienții lor adevărați sunt cei care-i plătesc pentru a primi acces la cunoștințele acumulate despre noi, pentru a ne putea convinge să le cumpărăm produsele sau pentru a ne influența în altă privință. O metodă de a evita

monopolul asupra datelor este să împărțim titlul de proprietate asupra datelor unor firme care le utilizează. Persoanele individuale ar deține și controla astfel datele lor cu caracter personal (un model care ar încuraja competiția, devreme ce oamenii ar fi liberi să-și mute datele la compania care oferă servicii mai bune). În cele din urmă, folosirea abuzivă a datelor nu este limitată la corporații: în regimurile totalitare sau chiar și în cele democrate cu numele, guvernele știu lucruri despre cetățenii lor pe care nici Orwell nu și le-ar fi putut imagina. Modul în care utilizează aceste informații nu este întotdeauna transparent și nu poate fi contracarat mereu.

Ideea de inteligență artificială în scopuri militare este înspăimântătoare. Ne putem imagina sisteme inteligente care sunt proiectate să acționeze autonom în baza datelor obținute în timp real, care pot acționa mult mai repede decât inamicul, inițiind războaie catastrofale. Astfel de războaie nu sunt neapărat convenționale sau nucleare. Dată fiind importanța rețelelor de computere în societatea modernă, este mult mai probabil ca războaiele inteligențelor artificiale să fie purtate în spațiul cibernetic. Consecințele pot fi la fel de grave.

În ciuda acestei pierderi a controlului, continuăm să mărșăluim inexorabil într-o lume în care inteligența artificială va fi peste tot. Indivizii nu vor putea rezista comodității și puterii, iar corporațiile și guvernele nu vor putea rezista avantajelor sale competitive. Dar iau naștere întrebări importante legate de viitorul muncii. Computerele sunt responsabile de pierderi considerabile în rândul meseriilor de muncă manuală în ultimele decenii, dar, până de curând, meseriile de birou – meserii pe care „le pot face doar oamenii" – erau considerate sigure. Brusc, acest lucru nu mai este adevărat. Contabili, mulți profesioniști din domeniul dreptului sau medicinei, analiști financiari și brokeri, agenți de turism – de fapt, mulți angajați la birou – vor dispărea ca urmare a programelor sofisticate de învățare automată. Vom face față unui viitor în care fabricile vor scoate pe bandă rulantă produse cu foarte

puțini angajați, iar transportul bunurilor este în mare parte automatizat, la fel ca multe alte servicii. Ce mai rămâne de făcut pentru oameni?

În 1930 – cu mult timp înainte de apariția computerelor, cu atât mai mult a inteligenței artificiale –, John Maynard Keynes scria, într-un eseu numit „Posibilitățile economice ale nepoților noștri", că, din cauza îmbunătățirilor productivității, societatea poate produce toate lucrurile de care are nevoie cu o săptămână de lucru de cincisprezece ore. De asemenea, a preconizat, odată cu creșterea timpului liber petrecut în mod creativ, sfârșitul banilor și avuției ca scop în viață:

> Vom putea să ne permitem să îndrăznim să evaluăm motivul banilor la valoarea sa adevărată. Dragostea față de bani ca posesie – diferită de dragostea față de bani ca mijloc de obținere a plăcerilor și realităților vieții – va fi recunoscută pentru ceea ce este, o morbiditate oarecum dezgustătoare, una dintre acele înclinații semicriminale, semipatologice pe care le predăm ridicând din umeri specialiștilor în boli mintale.

Din păcate, predicția lui Keynes nu s-a adeverit. Deși productivitatea a crescut, într-adevăr, sistemul – posibil inerent într-o economie de piață – nu a avut ca efect mai puține ore de lucru. În schimb, s-a întâmplat ceea ce antropologul și anarhistul David Graeber a descris ca fiind creșterea „meseriilor de rahat"*. În timp ce meseriile care produc elemente esențiale precum alimente, adăpost și bunuri au fost în mare parte automatizate, am văzut o extindere enormă a unor sectoare precum legislația corporatistă, administrarea academică și a sănătății (în opoziție cu predatul efectiv, cercetarea și practica medicinei), „resursele umane" și relațiile publice, fără a mai menționa industriile noi precum serviciile financiare și telemarketingul, și industriile auxiliare în așa

* https://strikemag.org/bullshit-jobs (n.a.).

numita economie a gigurilor, care servește persoanelor care sunt prea ocupate făcând toată munca suplimentară.

Cum se vor descurca societățile cu tehnologia care distruge rapid întregi profesii și lasă un număr mare de oameni fără loc de muncă? Unii argumentează că această grijă se bazează pe o premisă falsă, pentru că iau naștere meserii care nu existau înainte, dar, subliniază Graeber, aceste meserii noi nu vor fi neapărat satisfăcătoare și nu ne vor face să ne simțim împliniți. În timpul primei revoluții industriale, a durat aproape un secol înainte ca mulți oameni să aibă o situație mai bună. Acea revoluție a fost posibilă doar pentru că guvernele vremii favorizau acerb drepturile de proprietate asupra muncii și cei mai mulți oameni (și toate femeile) nu aveau dreptul de vot. În societățile democratice de astăzi, nu este clar dacă populația va tolera o răsturnare atât de spectaculoasă a societății în baza promisiunii că, „în cele din urmă", lucrurile se vor îmbunătăți.

Chiar și această viziune edulcorată va depinde de reformarea radicală a educației și a procesului de învățare de-a lungul vieții. Revoluția industrială a determinat schimbări sociale enorme de acest tip, inclusiv o alunecare către educația universală. Dar nu se va întâmpla decât dacă o facem noi să se întâmple: esențialmente, este vorba despre putere, mijlocire și control. Ce urmează pentru, să spunem, șoferul de taxi sau de camion, în vârstă de patruzeci și șase de ani, în era mașinilor autonome?

O idee care a fost vehiculată este cea a venitului de bază universal, care va permite cetățenilor să-și urmărească interesele, să se specializeze pentru alte ocupații și, în general, să fie liberi să trăiască o viață decentă. Însă economiile de piață, care au la bază o cerere din ce în ce mai mare din partea consumatorului mai mult decât orice altceva, probabil nu vor tolera această inovație. De asemenea, există senzația că a găsi sens în munca prestată este un element esențial pentru demnitatea umană și sentimentul de împlinire. Deci altă posibilitate ar fi ca bogăția enormă generată de productivitatea crescută datorită automatizării să fie redistribuită pentru meseriile ce necesită muncă și creativitate umane, în

domenii precum arta, muzica, asistența socială și alte activități ce merită efortul. În cele din urmă, ce meserii vor fi satisfăcătoare sau productive și care vor fi „de rahat" ține de părerea fiecăruia și pot varia de la societate la societate, precum și în funcție de timp.

Până acum m-am concentrat pe consecințele practice ale inteligenței artificiale. Ca om de știință, ce mă deranjează pe mine este posibilitatea de a pierde înțelegerea. Acum acumulăm date la o rată incredibilă. Doar în laboratorul meu, un experiment generează mai mult de un terabyte de date pe zi. Aceste date sunt transformate în mesaje, analizate și reduse până când ajungem la un rezultat interpretabil. Dar în toată această analiză de date, credem că știm ce se întâmplă. Știm ce fac programele, pentru că noi suntem cei care am proiectat algoritmii de la zero. Prin urmare, când computerele noastre generează un rezultat, simțim că îl înțelegem intelectual.

Noile programe de învățare automată sunt diferite. Recunoscând tiparele prin intermediul rețelelor neuronale profunde, generează concluzii și nu știm exact cum. Când descoperă relații, noi nu le înțelegem în același fel ca și cum am fi dedus acele relații noi înșine, utilizând un cadru teoretic subiacent. Pe măsură ce seturile de date devin mai mari, nu le vom putea analiza noi înșine, chiar și cu ajutorul computerelor; mai degrabă ne vom bizui în întregime pe computere să facă analiza pentru noi. Așadar, dacă cineva ne întreabă cum știm ceva, vom spune pur și simplu că mașina a analizat datele și a produs concluzia.

Într-o bună zi, un computer poate prezenta un rezultat cu totul nou – de exemplu, o teoremă matematică a cărei demonstrație sau chiar a cărei ipoteză nu poate fi înțeleasă de niciun om. Acesta este un mod filosofic diferit de modul în care am practicat știința până acum. Sau, cel puțin, de modul în care am crezut că am practicat-o. Unii ar putea argumenta că nu știm nici cum ajung creierele noastre la anumite concluzii și că aceste metode sunt o cale de a imita învățarea realizată de creierul uman. Cu

toate acestea, mie mi se pare înspăimântătoare această posibilă pierdere a înțelegerii.

În ciuda avansurilor remarcabile din domeniul tehnicii de calcul, febra inteligenței artificiale generale – o mașină generală de inteligență care va gândi ca un om și, probabil, va dezvolta conștiință – mie îmi miroase a science-fiction, în mare parte pentru că nu înțelegem creierul la acel nivel de detaliu. Nu doar că nu înțelegem ce este conștiința, nu înțelegem nici măcar o problemă relativ simplă, de exemplu, cum reușim să ținem minte un număr de telefon. În această întrebare simplă, trebuie să luăm în considerare o mulțime de factori. Cum știm că este un număr? Cum îl asociem unei persoane, unui nume, unei fețe și altor caracteristici? Chiar și astfel de întrebări aparent triviale implică totul de la cogniția de nivel superior și memorie până la modul în care o celulă stochează informațiile și cum interacționează neuronii.

În plus, aceasta este doar o sarcină printre multe altele pe care creierul le face fără efort. În timp ce mașinile vor face, fără îndoială, lucruri din ce în ce mai minunate, este puțin probabil să înlocuiască gândirea, creativitatea și viziunea umane. Eric Schmidt, fostul președinte al companiei Google, a spus într-un interviu recent la Muzeul de Științe din Londra că până și simpla desemnare a unui robot să curețe masa, să spele vasele și să le așeze va fi o provocare uriașă. Calculele implicate de a programa toate mișcările pe care trebuie să le facă corpul pentru a arunca o minge cu precizie sau de a face slalom pe schiuri sunt vaste. Creierul poate face toate acestea, făcând în același timp și calcule matematice sau muzică și *inventând* jocuri precum șah și Go, nu doar să le joace. Avem tendința să subestimăm complexitatea și creativitatea creierului uman și cât de uimitor de generalist este.

Dacă inteligența artificială urmează să devină mai umană în abilitățile sale, comunitățile de învățare automată și de neuroștiințe trebuie să interacționeze îndeaproape, ceva ce se întâmplă deja. Câteva dintre numele cele mai mari din prezent pentru inteligența artificială – cum ar fi Geoffrey Hinton, Zoubin Ghahramani

și Demis Hassabis – au experiență în neuroștiința cognitivă și succesul lor s-a datorat cel puțin parțial încercărilor de a modela comportamente asemănătoare creierului în algoritmii lor. În același timp, a înflorit și neurobiologia. Au fost dezvoltate tot soiul de instrumente pentru a vedea ce neuroni se activează și pentru a-i manipula genetic și a vedea ce se întâmplă în timp real cu intrările. Mai multe țări au lansat inițiative de neuroștiință cu țeluri foarte înalte pentru a vedea dacă putem să dezvăluim misterele creierului. Progresele în domeniul inteligenței artificiale și neuroștiinței par să meargă mână în mână; fiecare domeniu îl poate avansa pe celălalt.

Mulți cercetători evoluționiști și filosofi precum Daniel Dennett au scos în evidență faptul că creierul uman este rezultatul a miliarde de ani de evoluție*. Inteligența umană nu este caracteristica specială pe care o credem noi a fi, ci doar un alt mecanism de supraviețuire, nu foarte diferit de sistemele noastre digestive sau imunitare, și acestea din urmă fiind ambele incredibil de complexe. Inteligența a evoluat pentru că ne-a permis să înțelegem lumea din jurul nostru, să planificăm pentru viitor și să facem față astfel unei game variate de evenimente pentru a supraviețui. Însă, cum scria Descartes, noi, oamenii, ne definim existența prin abilitatea noastră de a gândi. Deci nu este surprinzător că, în mod antropomorfic, temerile noastre ce vizează inteligența artificială reflectă această credință că inteligența noastră este cea care ne face speciali.

Dar, dacă facem un pas înapoi și privim viața de pe Pământ, observăm că suntem departe de a fi cea mai rezistentă specie. Dacă vom fi cuceriți la un moment dat, vom fi cuceriți de cele mai vechi forme de viață de pe Pământ, cum ar fi bacteriile, care pot trăi oriunde, din Antarctica până la vulcanii submarini de pe platforma oceanică, zone mai fierbinți decât apa clocotită, sau în medii acide care ne-ar topi pe mine și pe tine. Așadar, când

* Vezi, de exemplu, cartea lui Dennett, *From Bacteria to Bach and Back: The Evolution of Minds* (New York: W.W. Norton, 2017) (n.a.).

oamenii întreabă unde ne îndreptăm, trebuie să punem întrebarea într-un context mai larg. Nu știu ce fel de viitor va inaugura inteligența artificială: dacă îi va face pe oameni sclavii ei sau niște ființe depășite sau dacă va fi o extindere utilă și binevenită a capacităților noastre, care ne va îmbogăți viețile. Dar sunt aproape sigur că computerele nu vor fi niciodată stăpânii bacteriilor.

Capitolul 19
Strategia umană

Alex „Sandy" Pentland

Alex „Sandy" Pentland este profesor Toshiba și profesor de științe și arte media la MIT; director al laboratoarelor de Dinamică Umană și Științe ale Conectării și al programului de Antreprenoriat Media Lab și autorul cărții *Social Physics*.

Alex „Sandy" Pentland, un exponent al ceea ce el a denumit „fizica socială", este interesat de construirea unor ecologii puternice între om și inteligența artificială. Se preocupă în același timp de potențialele pericole ale sistemelor de luare a deciziilor în care datele preiau controlul efectiv asupra creativității umane și creativitatea umană este surghiunită în fundal.

Apariția Big Data, consideră el, ne-a conferit ocazia de a ne reinventa civilizația: „Acum putem începe să privim cu adevărat detaliile interacțiunii sociale și ce efecte au acestea, nemaifiind limitați la valori medii, cum ar fi indicii de piață sau rezultatele alegerilor. Aceasta este o modificare uluitoare. Abilitatea de a vedea detaliile pieței, ale revoluțiilor politice și de a le putea preconiza și controla este cu

siguranță un caz de foc prometeic – poate fi utilizat pentru bine sau pentru rău. Big Data ne-a dus în niște vremuri interesante".

La întâlnirea noastră de grup din Washington, Connecticut, a mărturisit că, citind despre conceptul de *feedback* al lui Norbert Wiener, „am simțit că-mi citesc propriile gânduri".

„După Wiener, oamenii au descoperit ori s-au concentrat pe faptul că există sisteme cu adevărat haotice, care pur și simplu nu sunt predictibile", spunea el, „dar, dacă analizăm sistemele umane socioeconomice, există un procentaj mare de variație care poate fi contabilizat și preconizat... Astăzi avem date de la tot felul de dispozitive digitale și de la toate tranzacțiile noastre. Faptul că totul este câmp de date înseamnă că poți măsura lucrurile în timp real în majoritatea aspectelor vieții umane și, din ce în ce mai mult, în toate aspectele vieții omenești. Faptul că avem computere și tehnici de învățare automată interesante înseamnă că putem construi modele predictive pentru sistemele umane în moduri în care nu se putea face acest lucru până acum."

În ultima jumătate de secol, ideea inteligenței artificiale și a roboților inteligenți a dominat gândirea cu privire la relația dintre oameni și computere. Pe de o parte, acest lucru se datorează faptului că este ușor să spui povești despre inteligența artificială și roboți, iar pe de alta, datorită succeselor ei timpurii (de exemplu, demonstratorii de teoreme ce au reprodus majoritatea *principiilor matematice* ale lui Whitehead și Russell) și finanțării militare masive. Viziunea mai veche și mai largă asupra cibernetici, care considera artificialul ca făcând parte dintre sistemele mai mari de *feedback* și influență reciprocă, s-a estompat din conștiința publică.

Însă în anii care au urmat viziunea cibernetică a crescut încet-încet și a preluat tacit controlul până la punctul în care este „în aer". Cercetarea de ultimă generație în majoritatea disciplinelor inginerești este încadrată acum în sisteme de *feedback* dinamice și acționate de fluxuri de energie. Chiar și inteligența artificială este reconstituită sub formă de sisteme „de consiliere" om/mașină și armata începe finanțarea la scară largă în acest domeniu – ceva ce probabil ar trebui să ne îngrijoreze mai mult decât dronele și roboții umanoizi independenți.

Dar, pe măsură ce știința și ingineria au adoptat o postură mai apropiată de cibernetică, a devenit clar că până și viziunea cibernetici este mult prea mică. Eram inițial centrați pe integrarea actorului individual, și nu pe proprietățile emergente ale unei rețele de actori. Nu este deloc surprinzător, deoarece matematica

rețelelor nu a existat până de curând, așa că o știință cantitativă a comportamentului rețelelor era imposibilă. Acum știm că studierea cazurilor individuale nu produce înțelegerea sistemului decât în anumite cazuri simple. Progresele recente în acest domeniu au fost prefigurate de înțelegerea faptului că „haosul" și, mai târziu, „complexitatea" reprezentau comportamentul tipic al sistemelor, dar acum putem merge mult dincolo de aceste înțelegeri statistice.

Începem să putem analiza, preconiza și chiar proiecta comportamentele emergente ale rețelelor eterogene complexe. Perspectiva cibernetică a agentului individual conectat poate fi acum extinsă pentru a acoperi sisteme complexe de persoane și mașini conectate, iar observațiile pe care le facem în această perspectivă mai largă sunt fundamental diferite de cele făcute din perspectiva cibernetică. Să ne gândim la rețea este analog cu a ne gândi la un întreg ecosistem. Cum ghidezi ecosistemele să crească în direcția potrivită? Ce înseamnă „direcția potrivită"? Astfel de întrebări sunt dincolo de granița gândirii cibernetice tradiționale.

Poate că cea mai izbitoare realizare este că oamenii au început deja să utilizeze inteligența artificială și învățarea automată pentru a ghida întregi ecosisteme, printre care se numără și multe ecosisteme cu oameni, creând astfel ecologii de oameni și inteligențe artificiale la un loc. Acum că totul devine „datificat", putem să măsurăm cele mai multe aspecte ale vieții omenești și, din ce în ce mai mult, aspecte ale întregii vieți. Acest lucru, alături de tehnicile de învățare automată noi și puternice, înseamnă că putem construi modele ale acestor ecologii altfel decât puteam până acum. Exemplele bine-cunoscute sunt modelele de prognoză meteo și de preconizare a traficului, care sunt extinse pentru a preconiza clima globală și pentru a planifica modernizarea și creșterea orașelor. Ingineria ajutată de inteligența artificială a ecologiilor este deja aici.

Dezvoltarea ecosistemelor om–inteligență artificială este, probabil, inevitabilă pentru o specie socială cum suntem noi. Am devenit sociali devreme în evoluția noastră, cu milioane de ani în urmă. Am început să facem schimb de informații pentru a

supraviețui, pentru a ne spori adaptarea la mediu. Am dezvoltat scrisul pentru a împărtăși idei abstracte și complexe, și, mai recent, am dezvoltat computerele pentru a ne spori capacitățile de comunicare. Acum dezvoltăm inteligența artificială și modele de învățare automată pentru ecosisteme, împărtășind predicțiile acestor modele pentru a ne configura împreună viitorul lumii în care trăim prin legi și acorduri internaționale noi.

Trăim un moment istoric fără precedent, în care disponibilitatea unor volume uriașe de date despre comportamentul uman și progresele învățării automate ne permit să abordăm probleme sociale complexe prin algoritmizarea procesului de luare a deciziilor. Oportunitățile unei astfel de ecologii om–inteligență artificială de a avea un impact social pozitiv prin decizii mai echitabile și mai transparente sunt evidente. Dar există și riscurile unei „tiranii a algoritmilor", în care experți în date nealeși conduc lumea. Alegerile pe care le facem acum sunt, probabil, și mai importante decât cele pe care le-am făcut în anii 1950, când au fost create inteligența artificială și cibernetica. Problemele par similare, dar nu sunt. Am avansat, și acum miza este mai mare. Nu mai este vorba doar despre roboți cu inteligență artificială versus persoane. Este vorba despre inteligența artificială care ghidează întregi ecologii.

Cum putem face un ecosistem om–inteligență artificială bun, ceva care să nu fie o societate a mașinilor, ci o cultură cibernetică în care putem trăi cu toții ca oameni – o cultură pe care să o resimțim ca omenească? Nu vrem să gândim la scară mică – de exemplu, să vorbim doar despre roboți și despre mașini autonome. Dorim să fie vorba despre o ecologie globală. Ne gândim să fie la scara Skynet. Dar cum am putea face Skynetul să fie de partea oamenilor?

Prima întrebare pe care trebuie să ne-o punem este: care este magia care face inteligența artificială prezentă să funcționeze? Unde greșește și unde nu?

Magia bună este că are ceva numit funcția de alocare a creditelor. Aceasta îți permite să iei „neuronii proști" – funcții liniare mărunte – și să-ți dai seama, într-o rețea mare, care dintre ei fac

munca efectivă și să-i întărești. Este o metodă de a face mai inteligente câteva comutatoare aleatorii puse în rețea dându-le *feedback* cu privire la ce funcționează și ce nu. Sună simplu, dar matematica pe care se bazează este complicată. Aceasta este magia care face inteligența artificială curentă să funcționeze.

Partea proastă este că, din cauza faptului că acei neuroni mici sunt proști, lucrurile pe care le învață nu se generalizează foarte bine. Dacă o inteligență artificială vede ceva ce nu a mai văzut sau dacă lumea se schimbă puțin, inteligența artificială este pasibilă de a face o greșeală uriașă. Nu are nicio idee despre context. Din anumite puncte de vedere, este cât se poate de departe de noțiunea originală a lui Norbert Wiener de cibernetică, pentru că nu este contextualizată; este un mic savant idiot.

Dar să ne imaginăm că îndepărtăm aceste limitări: să ne imaginăm că, în loc să folosim neuroni proști, folosim neuroni în care au fost integrate cunoștințe despre lumea reală. Poate că, în loc de neuroni liniari, folosim neuroni care sunt funcții în fizică și apoi încercăm să potrivim datele fizice. Sau poate că integrăm multe cunoștințe despre oameni și cum interacționează ei unul cu celălalt – statisticile și caracteristicile oamenilor.

Când adăugăm aceste cunoștințe contextuale și le conferim și o funcție bună de atribuire a creditelor, putem lua datele din observații și folosi funcția de atribuire a creditelor pentru a întări funcțiile care produc răspunsuri bune. Rezultatul este o inteligență artificială care funcționează foarte bine și poate generaliza. De exemplu, în rezolvarea problemelor fizice, adesea necesită doar câteva puncte de date zgomotoase pentru a obține ceva ce este o descriere frumoasă a unui fenomen, pentru că introducem cunoștințe despre cum funcționează fizica. Această perspectivă prezintă un contrast uriaș cu inteligența artificială normală, care necesită milioane de exemple pentru învățare și este foarte sensibilă la zgomot. Adăugând cunoștințele contextuale corespunzătoare obținem mult mai multă inteligență.

Similar cu cazul sistemelor fizice, dacă vom face neuroni care știu multe despre cum învață oamenii unul de la altul, atunci vom

putea detecta mode omenești și prognoza tendințele de comportament ale oamenilor surprinzător de precis și eficient. Această „fizică socială" funcționează, deoarece comportamentul uman este determinat în aceeași măsură de tiparele culturii noastre, cât și de gândirea individuală și rațională. Aceste tipare pot fi descrise matematic și utilizate pentru a face predicții precise.

Această idee de funcție de atribuire a creditelor care întărește conexiunile dintre neuronii care fac cea mai bună treabă este nucleul inteligenței artificiale curente. Dacă faci acei neuroni mici mai inteligenți, inteligența artificială devine mai inteligentă. Dar ce s-ar întâmpla dacă am înlocui neuronii cu oameni? Oamenii au multe capacități. Știu multe despre lume. Pot percepe lucrurile într-un mod în general competent și omenesc. Ce s-ar întâmpla dacă am avea o rețea de oameni în care am putea întări conexiunile care ajută și minimaliza conexiunile care nu ajută?

Începe să sune ca o societate sau o companie. Cu toții trăim într-o rețea socială de oameni. Suntem „întăriți" când facem lucruri care par să ajute pe toată lumea și descurajați să facem lucruri care nu sunt apreciate. Cultura este rezultatul acestui tip de inteligență artificială omenească aplicată problemelor omenești; este procesul de a construi structuri sociale întărind conexiunile bune și penalizându-le pe cele rele. După ce realizăm că putem folosi acest cadru general pentru inteligența artificială și putem crea o inteligență artificială umană, întrebarea devine: Care este modul corect de a face asta? Este o idee sigură? Este nebunească?

Eu și studenții mei analizăm modul în care oamenii iau decizii din baze de date uriașe de decizii financiare, decizii de afaceri și multe alte tipuri de decizii. Am constatat că oamenii iau adesea decizii într-un mod care imită algoritmii de atribuire a creditelor folosiți de inteligența artificială și care funcționează pentru a face comunitatea mai isteață. O caracteristică interesantă a acestei activități este că analizează o problemă clasică din evoluție cunoscută sub numele de problema selecției de grup. Nucleul acestei probleme este: cum putem selecta cultura în evoluție, când indivizii sunt cei care se reproduc? Avem nevoie de ceva ce selectează

cele mai bune culturi și cele mai bune grupuri, dar, în același timp, selectează și cei mai buni indivizi, pentru că ei reprezintă unitățile care transmit genele.

Când formulăm întrebarea astfel și parcurgem literatura matematică, descoperim că există o metodă foarte bună. Este numită „distribuția eșantionului" de tip Thompson, un algoritm matematic utilizat pentru alegerea, dintr-un set de acțiuni posibile cu rezultate necunoscute, acțiunea care maximizează recompensa așteptată de la acțiuni. Cheia este reprezentată de eșantionarea socială, un mod de a combina dovezile, de a explora și exploata în același timp. Are neobișnuita proprietate de a fi simultan cea mai bună strategie atât pentru individ, cât și pentru grup. Dacă folosim grupul ca bază de selecție și grupul este eliminat sau întărit, îi selectăm în același timp și pe indivizii de succes. Dacă-i selectăm pe indivizi și fiecare individ face ce este bine pentru sine, atunci face automat ce este bine pentru grup. Este o aliniere uimitoare a intereselor și utilităților și oferă observații importante în privința modului în care cultura se potrivește în selecția naturală.

Eșantionarea socială, explicată simplu, înseamnă să te uiți în jur la acțiunile oamenilor care îți seamănă, observând ce este popular și apoi copiind dacă ți se pare o idee bună. Propagarea ideilor are această funcție de popularitate la bază, dar adoptarea individuală ține și de felul în care ideea funcționează pentru individ – o atitudine reflexivă. Când combinăm eșantionarea socială și judecata personală, obținem un proces superior de luare a deciziilor. Este uimitor, pentru că acum avem o rețetă matematică de a face cu oamenii ce fac toate acele tehnici de inteligență artificială cu neuronii proști ai computerelor. Avem posibilitatea de a alătura oameni pentru a lua decizii mai bune, dată fiind experiența din ce în ce mai mare.

Așadar ce se întâmplă în lumea reală? De ce nu facem astfel tot timpul? Ei bine, oamenii sunt priceputi la asta, dar există posibilități de a o da în bară. Una este prin publicitate, propagandă sau „știri false". Există multe moduri de a-i face pe oameni să creadă că ceva este popular atunci când nu este, iar acest lucru

distruge utilitatea eșantionării sociale. Modul în care putem crea grupuri de oameni mai inteligenți, modul în care putem face o inteligență artificială umană va funcționa doar dacă le putem transmite un *feedback* util. Dacă acțiunile fiecărei persoane au funcționat pentru ea sau nu, e ceea ce ar trebui să stea la baza acestui proces.

Aceasta este și cheia mecanismelor de inteligență artificială. Ce fac este să analizeze dacă au executat sarcinile corect. Dacă da, plus unu; dacă nu, minus unu. Avem nevoie de acel *feedback* sincer pentru a face acest mecanism omenesc să funcționeze bine, și avem nevoie de feluri bune de a ști ce fac alți oameni pentru a putea evalua corect popularitatea și probabilitatea de a fi o alegere bună.

Următoarea etapă este construirea acestei funcții de atribuire a creditelor, această funcție de *feedback*, pentru oameni, pentru a putea face un ecosistem bun pentru om și artificial – o organizație inteligentă și o cultură inteligentă. Într-un fel, trebuie să duplicăm câteva observații preliminare care au condus la crearea recensământului american, de exemplu – să încercăm să găsim dovezi de bază cu care toată lumea să fie de acord și să înțelegem astfel că transmiterea cunoștinței și culturii poate avea loc într-un mod veritabil și eșantionarea socială poate funcționa eficient.

Putem aborda problema creării unei funcții de atribuire a creditelor în multe contexte. În companii, de exemplu, poate fi făcută cu ecusoane de identificare digitală care descoperă cine este conectat cu cine, pentru a putea evalua tiparul de legături în raport cu rezultatul companiei, zilnic sau săptămânal. Funcția de atribuire a creditelor se întreabă dacă acele legături au ajutat la rezolvarea problemelor sau la inventarea unor soluții noi, și întărește legăturile utile. Când poți obține acel *feedback* cantitativ – ceea ce este dificil, întrucât cele mai multe lucruri nu se măsoară cantitativ –, atât productivitatea, cât și rata de inovare din organizație pot fi îmbunătățite semnificativ. Aceasta este, de exemplu, baza metodei de „îmbunătățire continuă" de la Toyota.

O etapă următoare ar fi să încercăm să facem același lucru, însă la scară, ceva ce eu numesc construirea unei rețele de

încredere pentru date. Poate fi considerat un sistem distribuit, ca internetul, dar cu capacitatea de a măsura cantitativ și de a comunica astfel calitățile societății umane, în același mod în care recensământul american face o treabă foarte bună informându-ne cu privire la populație și speranța de viață. Deja implementăm exemple prototip de rețele de încredere la scară în mai multe țări, bazate pe datele și standardele de măsurare expuse în obiectivele de dezvoltare sustenabilă ale Organizației Națiunilor Unite.

La orizont avem o viziune pentru cum am putea face umanitatea mai inteligentă construind o inteligență artificială umană. Este o viziune compusă din două părți. Una este reprezentată de datele în care putem avea cu toții încredere – date care au fost aprobate de o comunitate largă, date pentru care algoritmii sunt cunoscuți și monitorizați, asemenea datelor recensământului pe care ne bazăm cu toții automat ca fiind cel puțin aproximativ corecte. Cealaltă este o evaluare onestă, bazată pe date, despre condițiile curente. Această a doua parte depinde de disponibilitatea datelor de încredere, prin urmare abia începe să fie dezvoltată. Datele de încredere și evaluările bazate pe date ale normelor, politicilor și guvernelor creează împreună o funcție de atribuire a creditelor care îmbunătățește sănătatea generală și inteligența societății.

Exact aici, știrile false, propaganda și reclamele se pun în calea înfăptuirii unei inteligențe mai mari a societății. Din fericire, rețelele de încredere ne permit să mergem înainte și să construim o societate mai rezistentă la problema camerei de rezonanță, aceste mode, aceste exerciții de nebunie. Am început să dezvoltăm o modalitate nouă de a stabili măsurile sociale pentru a ne putea ajuta să vindecăm câteva afecțiuni de care suferă societatea de astăzi. Folosim date deschise din toate sursele, încurajând o reprezentare corectă a lucrurilor pe care le aleg oamenii, într-un cadru matematic îngrijit care poate amuți ecourile și încercările de manipulare a noastră.

Despre polarizare și inegalitate

Polarizarea extremă și segregarea după venit sunt prezente aproape în orice colț al lumii și amenință să destrame guvernele și societățile civile. Din ce în ce mai mult, mass-media devine un traficant de adrenalină împins de la spate de clicurile date pe reclame și nu mai reușește să raporteze echilibrat evenimentele, într-un discurs rațional, iar degradarea mass-mediei îi năucește pe oameni. Nu mai știu ce să creadă, putând fi astfel ușor manipulați. Există o nevoie reală de a înrădăcina diversele noastre culturi în standarde de încredere și care au la bază date cu care suntem cu toții de acord și de a putea să știm ce comportamente și politici funcționează sau nu.

În convertirea la o societate digitală, am pierdut contactul cu noțiunile tradiționale de adevăr și dreptate. Dreptatea era înainte preponderent informală și normativă. Acum am formalizat-o. În același timp, aceasta nu mai este la îndemâna celor mai mulți oameni. Sistemele noastre legale eșuează cum nu eșuau înainte, fix pentru că sunt acum mai formale, mai digitale, mai puțin încorporate în societate.

Ideile despre dreptate sunt foarte diferite în jurul lumii. Unul dintre elementele diferențiatoare de bază este următorul: tu sau părinții tăi vă amintiți când răufăcătorii au venit cu arme și v-au luat totul? Dacă îți aduci aminte, atunci atitudinea ta față de dreptate este diferită de cea a cititorului obișnuit al acestui eseu. Provii din pătura superioară a societății? Sau ești cineva care a văzut canalele dinăuntru? Perspectiva ta asupra dreptății depinde de viața pe care ai trăit-o.

Un test obișnuit pe care-l am pentru cetățenii americani este următorul: cunoști pe cineva care are o camionetă? Este cel mai bine vândut autovehicul din Statele Unite și, dacă nu știi astfel de oameni, nu ești în contact cu mai mult de 50% din populația Americii. Segregarea fizică stă la baza segregării conceptuale. Cei mai mulți americani privesc dreptatea, accesul și corectitudinea foarte diferit de locuitorul obișnuit din Manhattan, să spunem.

Dacă privim tiparele de mobilitate – unde merg oamenii – dintr-un oraș obișnuit, observăm că oamenii din cincimea superioară (familiile care au locuri de muncă la birouri) și cincimea inferioară (oamenii care trăiesc din șomaj sau ajutor social) nu comunică aproape niciodată. Nu merg în aceleași locuri, nu vorbesc despre aceleași lucruri. Locuiesc cu toții în același oraș, cu numele, dar este ca și cum ar fi două orașe diferite – și aceasta este probabil cea mai importantă cauză a plăgii polarizării din prezent.

Despre bogăția extremă

Aproximativ două sute dintre cei mai bogați oameni din lume au promis că vor da mai mult de jumătate din averea lor fie în timpul vieții, fie prin testament, creând astfel o pluralitate de voci în spațiul fundațiilor*. Bill Gates este probabil cel mai cunoscut exemplu. A decis că, dacă guvernul nu o face, o va face el. Vrei plase de țânțari? Le face el. Vrei medicamente antivirale? Le face el. Facem ca diverse persoane interesate să acționeze ca fundații dedicate binelui public, însă acestea au versiuni diferite pentru ceea ce consideră a fi binele public. Această diversitate de obiective a creat o mare parte din frumusețea lumii contemporane. Acțiuni întreprinse în afara guvernelor de organizații precum Fundația Ford și Fundația Sloan, care pariază pe lucruri pe care nimeni altcineva nu ar paria, au schimbat lumea în bine.

Bineînțeles, acești miliardari sunt oameni, cu metehne omenești, și nu totul este neapărat așa cum ar trebui să fie. Pe de altă parte, aceeași situație a existat și când au început să se construiască căi ferate. Unii oameni au făcut averi uriașe. Mulți au dat faliment. Noi, oamenii obișnuiți, ne-am ales cu căile ferate. Asta e bine. La fel și cu energia electrică; la fel cu multe tehnologii noi. Există un proces de agitație care aruncă pe cineva sus și, mai târziu, îi duce pe el sau pe moștenitorii săi jos. Bulele de bogăție extremă erau o caracteristică a sfârșitului secolului al XIX-lea și

* https://givingpledge.org/About.aspx (n.a.).

începutului secolului XX, când au fost inventate motoarele cu abur, căile ferate și luminile electrice. Averile pe care le-au creat dispăreau complet în două–trei generații.

Dacă Statele Unite ar fi fost ca Europa, mi-aș fi făcut probleme. Ce constatăm în Europa este că aceleași familii au ținut de averi sute de ani, deci sunt înrădăcinate nu doar ca avuție, dar și în sistemul politic și în alte domenii. Dar, până acum, SUA a evitat acest fel de sistem ereditar de clasă. Bogăția extremă nu a rămas de neatins, ceea ce este bine. Nu trebuie să fie de neatins. În cazul în care câștigi la loterie, primești miliardul de dolari, dar nepoții tăi trebuie să muncească pentru a trăi.

Despre inteligența artificială și societate

Oamenii se tem de inteligența artificială. Poate că așa și trebuie. Dar e bine să realizeze că inteligența artificială se hrănește cu date. Fără date, inteligența artificială nu reprezintă nimic. Nu trebuie să fii cu ochii pe inteligența artificială; în schimb trebuie să fii cu ochii pe ce consumă și ce face. Cadrul rețelei de încredere pe care l-am implementat, cu ajutorul unor națiuni din Uniunea Europeană și din alte zone, este unul în care putem avea algoritmii noștri, putem avea inteligența noastră artificială, dar putem să vedem ce a intrat și ce a ieșit din proces pentru a ne putea întreba: este aceasta o decizie discriminatorie? Asta vrem ca oameni? Sau este ceva puțin cam bizar?

Cea mai grăitoare analogie este că agențiile de reglementare, birocrațiile și părțile din guvern se aseamănă foarte mult cu inteligența artificială: preiau regulile pe care le numesc legi sau reglementări și adaugă la ele date guvernamentale, apoi iau decizii care ne afectează viețile. Partea care este rea din sistemul curent este că avem foarte puțin control asupra acestor departamente, agenții de reglementare și birocrații. Singurul control pe care-l avem este votul – oportunitatea de a alege o altă persoană. Trebuie să facem controlul asupra birocraților mult mai fin. Trebuie să înregistrăm datele care au intrat în fiecare decizie și să analizăm rezultate în

funcție de diversele părți interesate – așa cum se intenționa inițial să facă legislaturile alese.

Dacă avem datele care intră și ies din fiecare decizie, putem să ne întrebăm cu ușurință: este acesta un algoritm bun? Face această inteligență artificială lucruri pe care noi, oamenii, le considerăm etice? Această abordare cu oameni în ciclul de funcționare este numită „algoritm deschis"; adică putem vedea ce folosesc inteligențele artificiale ca intrări și ce decid folosind acele date de intrare. Dacă vedem aceste două lucruri, vom ști dacă facem bine sau nu. Se pare că nu este greu de făcut. Dacă vom controla datele, vom controla inteligența artificială.

Un lucru pe care adesea oamenii omit să-l menționeze este că toate grijile legate de inteligența artificială sunt identice cu grijile legate de guvernele contemporane. În cea mai mare parte a proceselor guvernamentale – sistemul de justiție etc. – nu există date fiabile despre ce fac și în ce situații anume. Cum putem ști dacă instanțele de judecată sunt imparțiale dacă nu știm intrările și ieșirile? Aceeași problemă ia naștere în legătură cu sistemele de inteligență artificială și poate fi abordată în același fel. Avem nevoie de date de încredere pentru a trage la răspundere guvernele curente cu privire la datele folosite și deciziile luate în baza acestora, iar inteligența artificială nu trebuie să se sustragă de la această regulă.

Inteligența generală din generația viitoare

Algoritmii prezenți de învățare automată folosiți de inteligențele artificiale sunt, la bază, simpli ca bună ziua. Funcționează, dar funcționează prin forță brută, deci au nevoie de sute de milioane de eșantioane. Funcționează pentru că poți aproxima orice cu multe bucăți simple. Aceasta este o observație-cheie a cercetării prezente din domeniul inteligenței artificiale – și anume, dacă folosești învățarea prin întărire pentru *feedback*-ul de atribuire a creditelor, poți face acele bucățele să aproximeze orice funcție arbitrară dorești.

Dar folosirea unor funcții greșite pentru a lua decizii înseamnă că abilitatea inteligenței artificiale de a lua decizii bune nu va fi generalizată. Dacă-i dăm inteligenței artificiale intrări noi, diferite, poate lua decizii complet nerezonabile. Sau, dacă situația se schimbă, atunci trebuie să o reinstruim. Există tehnici amuzante pentru a găsi „spațiile nule" în aceste sisteme de inteligență artificială. Acestea sunt intrări despre care inteligența artificială crede că sunt exemple valabile pentru ceea ce a fost antrenată să recunoască (de exemplu fețe, pisici etc.), dar pentru un om sunt exemple nebunești.

Inteligența artificială curentă face statistici descriptive într-un mod care nu este științific și ar fi aproape imposibil de transformat în știință. Pentru a construi sisteme robuste, trebuie să cunoaștem știința din spatele datelor. Sistemele pe care le văd ca inteligențe artificiale de nouă generație iau naștere din abordarea bazată pe știință: dacă urmează să creezi o inteligență artificială care să rezolve o problemă fizică, atunci trebuie să integrezi în ea legile fizicii drept funcții descriptive, în locul acelor neuroni proști. De exemplu, știm că fizica folosește funcții polinomiale, sinusoidale și exponențiale, deci acestea ar trebui să fie funcțiile de bază, și nu mici neuroni liniari. Folosind aceste funcții de bază mai potrivite, avem nevoie de mult mai puține date, putem face față unui zgomot mult mai mare și obținem rezultate mult mai bune.

Ca în exemplul cu fizica, dacă dorim să construim o inteligență artificială care să lucreze cu comportamentul uman, trebuie să construim proprietățile statistice ale rețelelor omenești în algoritmi de învățare automată. Când înlocuim neuronii proști cu neuroni care conțin câteva elemente de bază despre comportamentul uman, putem identifica tendințe cu foarte puține date și putem face față unui nivel uriaș de zgomot.

Faptul că oamenii au o înțelegere „intuitivă" pe care o folosesc în majoritatea problemelor sugerează ceea ce eu numesc strategia umană: societatea umană este o rețea la fel ca rețelele neuronale instruite pentru învățare profundă, dar „neuronii" din societățile

omenești sunt mult mai inteligenți. Tu și eu avem puteri descriptive surprinzător de generale pe care le folosim pentru a înțelege o gamă variată de situații și putem recunoaște ce conexiuni trebuie întărite. Asta înseamnă că ne putem configura rețelele sociale să funcționeze mult mai bine și, posibil, să învingă toată acea inteligență artificială bazată pe mașini la propriul ei joc.

Capitolul 20
Transformarea invizibilului în vizibil: arta față în față cu inteligența artificială

Hans Ulrich Obrist

Hans Ulrich Obrist este director artistic la Galeria Serpentine din Londra și autorul cărților *Ways of Curating* și *Lives of the Artists, Lives of the Architects.*

„URGENT! URGENT!", țipa copia cc a unui e-mail, unul dintre zecile de e-mailuri pe care le-am primit când am pornit telefonul în zona de bagaje din aeroportul Malpensa după zborul lung de la JFK. „Vizionarul american John Brockman ajunge în această dimineață la Grand Hotel Milan. TREBUIE, repet, TREBUIE să-i faci o vizită." E-mailul era semnat HUO.

Cu o seară înainte, stând în sala de așteptare de la JFK, avusesem ideea strălucită să-i scriu prietenului și colaboratorului meu de mult timp, care locuiește în Londra, peripateticul curator de artă Hans Ulrich Obrist (pe care toată lumea îl știe ca HUO), și să-l întreb dacă există vreo persoană în Milano pe care ar trebui să o cunosc.

Când m-am instalat la hotel, telefonul a început să zbârnâie și o procesiune de artiști, designeri și arhitecți italieni au sunat să-mi solicite o întrevedere, printre care și

Enzo Mari, artistul modernist și designerul de mobilă; Alberto Garutti, ale cărui strategii estetice au inspirat un dialog între arta contemporană, spectator și spațiul public; și designerul de modă Miuccia Prada, care „vă solicită prezența la ceai în această după-amiază la sediul Prada". Și astfel, mulțumită lui HUO, „marele gânditor american vizionar", obosit după zbor, a trecut cu pași grei de prima zi a sa în Milano, în noiembrie 2011.

HUO este *sui generis*: trăiește o zi de douăzeci și patru de ore, dormind (cred) pe apucate și angajând asistenți cu normă întreagă, care lucrează ture de opt ore și îi sunt disponibili 24 de ore pe zi, 7 zile pe săptămână. Într-o perioadă recentă de doi ani a vizitat galerii de artă în China sau în India patruzeci de weekenduri în fiecare an – plecând din Londra joia și revenind la birou lunea. Anul trecut, din nou, ArtReview l-a clasificat din nou ca fiind numărul unu pe lista lor anuală „Power 100".

Recent, am colaborat într-o comisie în timpul evenimentului Serpentine „GUEST, GHOST, HOST: MACHINE!", care a avut loc în noul sediu al Primăriei Londrei. Ni s-au alăturat Venki Ramakrishnan, Jaan Tallinn și Andrew Blake, director de cercetare la Institutul Alan Turing. Evenimentul se subscria misiunii lui HUO de a aduce laolaltă arta și știința: „Curatorul nu mai este înțeles pur și simplu ca o persoană care umple un spațiu cu obiecte", spune el, „ci și ca persoana care pune în contact sfere culturale diferite, inventează noile caracteristici de afișare și face legăturile necesare care permit întâlniri și rezultate neașteptate".

În introducerea celei de-a doua ediții a cărții sale *Understanding Media*, Marshall McLuhan a observat capacitatea artei de a „anticipa progresele sociale și tehnologice ale viitorului". Arta este „un sistem de alarmă timpuriu" care ne arată noile dezvoltări ce vor urma și ne permite să „ne pregătim să le facem față... Arta ca mediu de radar preia funcția de instruire perceptuală indispensabilă".

În 1964, când a fost publicată pentru prima dată cartea lui McLuhan, artistul Nam June Paik tocmai își construia *Robot K-456* pentru a experimenta cu tehnologiile care vor urma să influențeze societatea. Lucrase și înainte cu televiziunea, provocând consumul pasiv al telespectatorului, și, mai târziu, a creat artă cu difuzări live prin satelit, utilizând noile medii mai puțin pentru divertisment și mai mult pentru a evidenția capacitățile sale poetice și interculturale (care sunt încă în mare parte neutilizate în prezent). Paik-ii vremurilor noastre, bineînțeles, lucrează acum cu internetul, imagini digitale și inteligență artificială. Lucrările și gândurile lor sunt din nou un sistem de alarmă timpuriu pentru dezvoltările care vor urma.

În calitate de custode, activitatea mea zilnică constă în a reuni opere de artă diverse pentru a conecta culturi diferite. De la începutul anilor 1990, organizez conversații și întâlniri cu practicanți ai unor diverse discipline, pentru a putea trece dincolo de reticența generală de acumulare a cunoștințelor. Deoarece eram interesat să

aud ce au de spus artiștii despre inteligența artificială, am organizat recent mai multe conversații între artiști și ingineri.

Motivul pentru care privesc cu interes inteligența artificială este că două dintre cele mai importante întrebări ale zilelor noastre sunt „Când de capabilă va deveni inteligența artificială?" și „Ce pericole se pot naște din ea?" Aplicările sale incipiente ne influențează deja viața de zi cu zi în moduri care sunt mai mult sau mai puțin recognoscibile. Există un impact în creștere asupra multor aspecte ale societății noastre, dar, dacă acest lucru este, în general, benefic sau malign este încă nesigur.

Mulți artiști contemporani urmăresc îndeaproape aceste dezvoltări. Articulează diverse îndoieli despre promisiunile inteligenței artificiale și ne amintesc să nu asociem termenul de „inteligență artificială" numai cu rezultate pozitive. La discuțiile curente despre inteligența artificială, artiștii contribuie cu perspectivele lor speciale și cu concentrarea pe întrebări legate de crearea imaginilor, creativitate și utilizarea programării ca instrument artistic.

Conexiunile profunde dintre știință și artă au fost observate deja de regretatul Heinz von Foerster, unul dintre arhitecții ciberneticii, care a lucrat cu Norber Wiener de la jumătatea anilor 1940, iar în anii 1960 a fondat domeniul ciberneticii secundare, în care observatorul este înțeles ca făcând parte din sistem, și nu ca o entitate externă. L-am cunoscut bine pe von Foerster și într-una dintre multele conversații pe care le-am purtat mi-a comunicat perspectiva sa asupra relației dintre artă și știință:

> Întotdeauna am perceput arta și știința ca pe niște domenii complementare. Nu trebuie să uităm că un om de știință este în anumite privințe și un artist. Inventează o tehnică nouă și o descrie. Folosește limbajul ca un poet sau ca autorul unui roman polițist și își descrie rezultatele. Din punctul meu de vedere, un om de știință trebuie să lucreze artistic dacă vrea să-și comunice cercetările. Este evident că vrea să le comunice și să vorbească cu alții. Un om de

știință inventează obiecte noi și problema este cum vor fi descrise. În toate aceste privințe, știința nu este foarte diferită de artă.

Când l-am întrebat cum definește cibernetica, von Foerster a răspuns:

> Substanța învățăturilor noastre trase din cibernetică este să gândim circular: A conduce la B, B la C, dar C se întoarce la A. Astfel de argumente nu sunt liniare, ci sunt circulare. Contribuția semnificativă a ciberneticii la gândirea noastră este acceptarea argumentelor circulare. Asta înseamnă că trebuie să privim procesele circulare și să înțelegem care sunt circumstanțele în care ia naștere un echilibru, deci o structură stabilă.

Astăzi, acolo unde algoritmii de inteligență artificială sunt aplicați la sarcinile zilnice, ne putem întreba cum este inclus factorul uman în aceste tipuri de procese și ce roluri pot juca arta și creativitatea în raport cu ei. Astfel există niveluri diferite de gândire despre explorarea relației dintre inteligență artificială și artă.

Prin urmare, ce au de spus artiștii contemporani despre inteligența artificială?

Prostia artificială

Hito Steyerl, o artistă care lucrează cu filme documentare și experimentale, ia în considerare două aspecte-cheie pe care trebuie să le avem în minte când ne gândim la consecințele inteligenței artificiale pentru societate. În primul rând, așteptările pentru așa-numita inteligență artificială, spune ea, sunt adesea exagerate, iar substantivul „inteligență" ne induce adesea în eroare; pentru a contracara acest efect, folosește termenul de „prostie artificială". În al doilea rând, scoate în evidență faptul că programatorii fac acum algoritmii software invizibili să fie vizibili prin

imagini, dar, pentru a înțelege și a interpreta mai bine aceste imagini, trebuie să aplicăm expertiza artiștilor.

Steyerl a lucrat cu computerele mulți ani și operele ei recente de artă au explorat tehnicile de supraveghere, roboții și jocuri pe calculator, precum în *How Not to be Seen* (2013), despre tehnologii de imagistică digitală, sau *HellYeahWeFuckDie* (2017), despre instruirea roboților în obiectivul echilibrului, care încă este dificilă pentru ei. Dar, pentru a explica noțiunea ei de prostie artificială, Steyerl face trimitere la un fenomen mai general, cum este utilizarea acum generalizată a „boților" Twitter, observând în conversația noastră:

> A fost și încă este o unealtă foarte populară în alegeri să angajezi armate Twitter să încline opinia publică și să devieze *hashtag*-urile populare etc. Aceasta este o inteligență artificială de un nivel foarte, foarte scăzut. Înseamnă două sau poate trei linii de script. Nu este deloc sofisticat. Însă implicațiile sociale ale acestui tip de prostie artificială, cum o numesc eu, sunt deja monumentale în politica globală.

După cum a fost observat pe scară largă, acest tip de tehnologie a fost văzut în multe postări Twitter automate înainte de alegerile prezidențiale din 2016, în Statele Unite, și, de asemenea, cu puțin timp înainte de votul Brexit. Dacă până și o tehnologie de inteligență artificială atât de simplă cum sunt acești „boți" influențează deja politica, ia naștere altă întrebare urgentă: „Cât de puternice vor fi tehnicile mult mai avansate din viitor?"

Vizibil/invizibil

Artistul Paul Klee vorbea adesea despre artă ca „a face invizibilul vizibil". În tehnologia informației, majoritatea algoritmilor lucrează invizibil, în fundal; rămân inaccesibili în sistemele pe care le folosim zi de zi. Dar, în ultima perioadă, s-a observat o revenire interesantă a vizualului în învățarea automată. Modurile

în care algoritmii de învățare automată ai inteligenței artificiale prelucrează datele au fost făcuți vizibili prin aplicații precum DeepDream de la Google, în care procesul de recunoaștere computerizată a tiparelor este vizualizat în timp real. Aplicația arată cum algoritmul încearcă să potrivească formele de animale cu orice intrare dată. Există multe alte programe de vizualizare bazate pe inteligență artificială care, în felul lor, fac și ele „invizibilul vizibil". Dificultatea din percepția publicului larg a acestor imagini este, după părerea lui Steyerl, că aceste tipare vizuale sunt privite fără critică drept reprezentări realiste și obiective ale procesului automat. Spune despre estetica unor astfel de vizualizări:

> Pentru mine, acest lucru dovedește că știința a devenit un subgen al istoriei artei... Avem acum multe tipare abstracte de computer care pot arăta ca o pictură de Paul Klee sau un Mark Rothko sau tot felul de alte abstracțiuni pe care le știm din istorie. Singura diferență, cred eu, este că, în gândirea științifică curentă, sunt percepute ca reprezentări ale realității, aproape ca imaginile unui documentar, în timp ce în istoria artei există o înțelegere foarte nuanțată a diverselor tipuri de abstractizări.

Ceea ce caută ea este o înțelegere mai profundă a imaginilor generate de computer și diversele forme estetice pe care le utilizează acestea. Evident, nu sunt generate cu scopul explicit de a respecta o anumită tradiției estetică. Inginerul IT Mike Tyka, într-o conversație purtată cu Steyerl, explica funcțiile acestor imagini:

> Sistemele de învățare profundă, în special cele vizuale, sunt de fapt inspirate de nevoia de a ști ce se întâmplă în cutia neagră. Scopul lor este să proiecteze acele procese înapoi în lumea reală.

Cu toate acestea, aceste imagini au implicații estetice și valori care trebuie luate în considerare. Se poate spune că, în timp ce

programatorii folosesc imagini pentru a ne ajuta să înțelegem mai bine algoritmii programelor, avem nevoie de cunoștințele artiștilor pentru a înțelege mai bine formele estetice ale inteligenței artificiale. Cum a evidențiat Steyerl, astfel de vizualizări sunt în general înțelese ca reprezentări „adevărate" ale proceselor, dar trebuie să fim atenți la esteticile lor corespunzătoare și la implicațiile acestora, care trebuie văzute critic și analitic.

În 2017, artistul Trevor Paglen a creat un proiect de a face vizibili acești algoritmi invizibili ai inteligenței artificiale. În *Sight Machine*, a filmat o reprezentație live a cvartetului Kronos și a procesat imaginile rezultate cu diverse programe software folosite pentru detectare facială, identificarea obiectelor și chiar ghidarea rachetelor balistice. A proiectat rezultatul acestor algoritmi, în timp real, înapoi pe ecrane în spatele scenei. Demonstrând cum interpretează performanța muzicienilor diversele programe, Paglen a arătat că algoritmii inteligenței artificiale sunt determinați întotdeauna de seturi de valori și interese pe care apoi le manifestă și reiterează, trebuind astfel interogați critic. Contrastul semnificativ dintre algoritmi și muzică ridică și problema relațiilor dintre percepția tehnică și cea omenească.

Computerele, ca instrument de creativitate, nu-l pot înlocui pe artist

Rachel Rose, o artistă video care se gândește la problemele puse de inteligența artificială, folosește tehnologia computerelor în crearea operelor sale. Filmele sale conferă privitorului o experiență a materialității prin intermediul imaginii în mișcare. Folosește colajul și stratificarea materialului pentru a manipula sunetul și imaginea, iar procesul de editare este probabil cel mai important aspect al activității sale.

De asemenea, vorbește despre importanța procesului de luare a deciziilor în activitatea sa. Pentru ea, procesul artistic nu respectă un tipar rațional. Într-o conversație pe care am avut-o,

împreună cu inginerul Kenric McDowell, de la Google Cultural Institute, a explicat această idee citând o povestioară din cartea regizorului de teatru Peter Brook *The Empty Space / Spațiul gol*, din 1968. Când Brook a conceput decorul pentru spectacolul său *Furtuna*, la sfârșitul anilor 1960, a început prin a face o grădină japoneză, dar apoi conceptul a evoluat, devenind o cutie albă, o cutie neagră, o scenografie realistă și așa mai departe. Și, în cele din urmă, s-a întors la ideea sa inițială. Brook scrie că a fost șocat de faptul că a pierdut o lună întreagă doar pentru a sfârși unde începuse. Dar acest lucru arată că procesul artistic creativ este o succesiune a cărei fiecare etapă se construiește pe cea anterioară și, în cele din urmă, ajunge la o concluzie impredictibilă, legată de propriul proces artistic de luare a deciziilor:

> Pentru mine, este foarte diferit de învățarea automată, pentru că la baza fiecărei decizii se află acest sentiment adânc care vine din ființa umană, care are de-a face cu empatia, care are de-a face cu comunicarea, care are de-a face cu întrebări cu privire la propria noastră mortalitate pe care doar un om le poate pune.

Această idee evidențiază diferența fundamentală dintre orice producție artistică umană și așa-zisa creativitate a computerelor. Rose vede inteligența artificială mai mult ca pe o cale de a crea instrumente mai bune pentru oameni:

> Îmi imaginez că învățarea automată îl poate ajuta pe artist nu în dezvoltarea unei subiectivități independente – cum ar fi compunerea unui poem sau crearea unei imagini –, ci în umplerea golurilor care au de-a face cu munca manuală, cum face Photoshopul cu diversele instrumente pe care le poți utiliza.

Și, deși aceste instrumente poate nu par spectaculoase, spune ea, „ar putea avea o influență importantă asupra artei", pentru că

pun la dispoziția artiștilor mai multe posibilități în activitatea lor creativă.

McDowell a adăugat că și el crede că există așteptări false de la inteligența artificială. „Am observat", spune el, „că ideea unui computer care să facă toate lucrurile pe care le facem și noi este oarecum magică". A continuat: „există aproape un fel de oglindă demonică în care privim și vrem ca imaginea să scrie un roman, să facă un film – vrem să predăm cumva această putere pe care o avem". În schimb, lucrează la proiecte în care oamenii colaborează cu mașinile. Unul dintre scopurile prezente ale cercetării din domeniul inteligenței artificiale este să găsească noi mijloace de interacțiune între oameni și software. Și arta, se poate spune, trebuie să joace un rol-cheie în această întreprindere, din moment ce se concentrează pe subiectivitatea noastră și pe aspecte omenești esențiale, cum ar fi empatia și mortalitatea.

Cibernetică/artă

Suzanne Treister este o artistă a cărei activitate dintre anii 2009 și 2011 servește drept exemplu pentru ceea ce se întâmplă la intersecția dintre tehnologiile noastre curente, arte și cibernetică. Treister a fost o pioniera în arta digitală încă din anii 1990, inventând, de exemplu, jocuri video imaginare și pictând capturi de ecran din ele. În proiectul său *HEXEN 2.0*, a analizat faimoasele conferințe de cibernetică Macy, care au fost organizate între 1946 și 1953, în New York, de ingineri și cercetători din domeniul științelor sociale, pentru a uni științele și pentru a dezvolta o teorie universală a mecanismelor minții.

În proiectul său, a creat treizeci de lucrări foto-text despre participanții la conferința sa (printre care se numărau și Wiener și von Foerster), a inventat cărți de tarot și a făcut un video bazat pe un montaj fotografic al unei „ședințe cibernetice de spiritism". Participanții la conferință se pot vedea stând în jurul unei mese rotunde, ca în ședințele de spiritism, în timp ce câteva dintre afirmațiile lor despre cibernetică se aud pe colajul audio – cunoștințe

raționale și superstiții combinate. De asemenea, a observat că anumiți participanți la ședință lucrau pentru armată; astfel, aplicarea ciberneticii putea fi privită ambivalent, chiar și atunci, ca o luptă între cunoștințele pure și utilizarea ei în controlul exercitat de stat.

Dacă privim lucrarea lui Treister despre participanții la conferința Macy, vedem că nu a fost inclus niciun artist vizual. Un dialog între artiști și oamenii de știință va fi fructuos în discuțiile viitoare și este puțin uimitor că nu și-au dat seama de acest lucru la momentul respectiv, dat fiind interesul profund al lui Foerster pentru artă. Și-a amintit într-una dintre conversațiile noastre cum relația lui cu domeniul își avea rădăcinile chiar în copilărie:

> Am crescut într-o familie de artiști. Adesea eram vizitați de poeți, filosofi, pictori și sculptori. Arta făcea parte din viața mea. Mai târziu, a început să mă intereseze fizica, întrucât se pare că aveam talent. Dar am rămas întotdeauna conștient de importanța artei pentru știință. Nu era o diferență mare pentru mine. Pentru mine, ambele aspecte ale vieții erau foarte asemănătoare și accesibile. Ar trebui să le vedem ca pe unul singur. Și un artist trebuie să reflecteze asupra lucrărilor sale. Trebuie să se gândească la gramatică și la limbaj. Un pictor trebuie să știe cum să-și gestioneze culorile. Gândiți-vă cât de intens erau cercetate culorile în timpul Renașterii. Voiau să știe cum un anumit pigment putea fi amestecat cu alții pentru a obține o anumită nuanță de roșu sau albastru. Chimiștii și pictorii colaborau îndeaproape. Cred că împărțirea artificială între științe și arte este greșită.

Deși pentru von Foerster relația dintre artă și știință a fost întotdeauna clară, pentru vremurile noastre această legătură nu este făcută încă. Există multe motive pentru a multiplica legăturile. Gândirea critică a artiștilor ar fi benefică în ceea ce privește pericolele inteligenței artificiale, deoarece ei ne atrag atenția

asupra întrebărilor pe care le consideră esențiale din perspectiva proprie. Cu progresele făcute de învățarea automată, artiștii au la dispoziție instrumente noi pentru activitatea lor. Și, pe măsură ce algoritmii inteligenței artificiale sunt făcuți vizibili prin imagini artificiale în moduri noi, expertiza și cunoștințele vizuale critice ale artiștilor vor fi folosite. Multe dintre aceste întrebări-cheie despre inteligența artificială sunt de natură filosofică și pot primi răspuns doar dintr-o perspectivă holistică. Modul în care prind viață în rândurile artiștilor aventuroși va merita urmărit.

Simularea unor lumi

În cea mai mare parte, lucrările artiștilor contemporani au fost meditații întruchipate ale impactului inteligenței artificiale asupra unor întrebări existențiale despre sine și interacțiunile noastre viitoare cu entitățile nonumane. Puțini însă au privit tehnologiile și inovațiile inteligenței artificiale ca pe materia primă a creațiilor lor și le-au sculptat în propria viziune. O excepție este reprezentată de artistul Ian Cheng, care a mers până în punctul în care a construit lumi întregi de ființe artificiale cu diverse grade de conștiință și inteligență. A numit aceste lumi „simulări live". Trilogia sa *Emisarii* (2015–2017) este plasată într-o lume ficțională postapocaliptică cu floră și faună, în care animale și creaturi controlate de inteligența artificială explorează peisajul și interacționează unele cu altele. Cheng folosește o grafică avansată, dar le-a programat cu multe erori și imperfecțiuni, creând astfel o atmosferă futuristă și anacronică în același timp. Prin trilogia sa, care cartografiază un istoric al conștiinței, pune întrebarea „ce este o simulare?"

În timp ce majoritatea lucrărilor artistice ce utilizează ultimele tehnologii din domeniul inteligenței artificiale se inspiră în special din învățarea automată, simulările live ale lui Chen o iau pe altă cale. Protagoniștii și poveștile care sunt îmbinate cu fiecare simulare episodică a *Emisarilor* folosesc sistemele logice și regulile complexe ale inteligenței artificiale. Ce este profund în scenele

sale care evoluează în mod continuu este că complexitatea ia naștere nu din dorințele/acțiunile unui singur actor sau zeu artificial, ci din constelațiile, coliziunile și evoluția constantă în simbioza dintre entități. Acest lucru dă naștere unor rezultate neașteptate și situații nesfârșite, care nu pot fi cunoscute – nu poți experimenta niciodată același moment dacă vezi repetat lucrările sale.

Cheng a susținut o conferință la maratonul Serpentine „GUEST, GHOST, HOST: MACHINE!", cu programatorul Richard Evans, care a proiectat recent Versu, o platformă bazată pe inteligență artificială pentru jocuri interactive de narare a unei povești. Lucrările lui Evans scot în evidență interacțiunea socială a personajelor din joc, care reacționează într-un spectru de comportamente posibile la alegerile făcute de jucătorii umani. În conversația lor, Evans a spus că un punct de plecare pentru proiectul său a fost faptul că primele jocuri de simulare, cum ar fi The Sims, nu au luat în considerare suficient importanța practicilor sociale. Protagoniștii simulați din jocuri se comportă adesea în moduri care nu corespund comportamentelor umane reale. Cunoașterea practicilor sociale limitează posibilitățile de acțiune, dar este necesară pentru a înțelege sensul acțiunilor noastre – ceea ce-l interesează pe Cheng pentru propriile sale simulări. Cu cât sunt mai bine determinați parametrii acțiunilor în anumite circumstanțe într-o simulare pe computer, cu atât este mai interesant pentru Cheng să experimenteze cu schimbări individuale și specifice. I-a spus lui Evans: „Consider că, în măsura în care am avea o inteligență artificială cu o capacitate mai bună de a răspunde contextelor sociale, dacă ajustăm o calitate obținem ceva artistic și frumos".

De asemenea, Cheng privește activitatea programatorilor și simulărilor cu inteligență artificială la bază drept crearea unor instrumente noi și sofisticate pentru experimentarea cu parametrii activităților noastre sociale de zi cu zi. În acest mod, implicarea artiștilor în domeniul inteligenței artificiale va conduce la tipuri noi de experimente deschise în artă. Astfel de posibilități

sunt – asemenea capacităților crescute ale inteligenței artificiale în general – încă în viitor. Recunoscând că aceasta este o tehnologie experimentală, incipientă, foarte departe de viziunile apocaliptice despre inteligențe artificiale superinteligente care vor cuceri lumea, Cheng își umple simulările cu avataruri prozaice, cum ar fi niște globule microbiene ciudate, câini și morți vii.

Astfel de discuții între artiști și ingineri, bineînțeles, nu sunt complet noi. În anii 1960, inginerul Billy Klüver a reunit artiști cu ingineri într-o serie de evenimente, iar în 1967 a pus bazele programului „Experimente în artă și tehnologie", cu Robert Rauschenberg și alții. În Londra, aproximativ în același timp, Barbara Steveni și John Latham, din Artist Placement Group, au dus lucrurile cu un pas mai departe, afirmând că ar trebui să existe artiști angajați la fiecare companie și în fiecare guvern. Astăzi aceste modele istorice pline de inspirație pot fi aplicate în domeniul inteligenței artificiale. Pe măsură ce inteligența artificială ajunge să fie implicată din ce în ce mai mult în viețile noastre de zi cu zi, crearea unui spațiu care este nondeterminist și nonutilitar în pluralitatea sa de perspective și diversitatea de înțelegeri va fi cu siguranță esențială.

Capitolul 21
Inteligențele artificiale versus copiii de patru ani

Alison Gopnik

Alison Gopnik este specialistă în psihologia dezvoltării la UC Berkeley. Printre cărțile sale se numără *The Philosophical Baby* și, mai recent, *The Gardener and the Carpenter: What the New Science of Child Development Tells Us About the Relationship Between Parents and Children.*

Alison Gopnik este lider internațional în domeniul învățării și dezvoltării copiilor și a fost una dintre fondatoarele domeniului „teoriei minții". A vorbit despre creierul de copil ca fiind un „puternic computer de învățare", poate din experiență personală. Copilăria ei din Philadelphia a fost un exercițiu în dezvoltarea intelectuală. „Alte familii își duceau copiii să vadă filmul *Sunetul muzicii* sau *Carusel*; noi am văzut *Fedra* lui Racine și *Sfârșitul jocului* de Samuel Beckett", își amintea ea. „Familia noastră citea cu voce tare romanul *Joseph Andrews*, scris de Henry Fielding în secolul al XVIII-lea, în jurul focului de tabără, când mergeam la camping."

Recent, a invocat modelele bayesiene ale învățării automate pentru a explica abilitatea excepțională a preșcolarilor de a trage concluzii despre lumea din jurul lor fără a beneficia de seturi de date eronate. „Cred că bebelușii și copiii sunt mai conștienți decât suntem noi, adulții", spunea ea. „Se pricep de minune să absoarbă multă informație din multe surse diferite deodată". I-a numit pe bebelușii și pe copiii tineri „departamentul de cercetare și dezvoltare al speciei umane". Asta nu înseamnă că este rece cu ei, ca și cum ar fi niște simple animale de laborator. Copiii par să se simtă foarte bine în compania ei și printre jucăriile luminoase și zgomotoase din laboratorul său de la Berkeley. Ani întregi după ce propriii săi copii au trecut de vârsta jucăriilor, ea le-a păstrat în birou.

Investigațiile sale în procesul nostru de învățare și paralelele făcute cu metodele de învățare profundă ale inteligenței artificiale continuă. „Se pare că este mult mai ușor să simulezi rațiunea unui adult cu multă educație decât să imiți învățarea obișnuită a unui bebeluș", spune ea. „Tehnica de calcul este încă cea mai bună – dacă nu singura – explicație științifică pe care o avem pentru modul în care un obiect fizic cum este creierul se poate purta inteligent. Dar, cel puțin deocamdată, nu avem aproape nicio idee cum este posibilă creativitatea pe care o vedem la copii."

Toată lumea a auzit despre ultimele progrese în domeniul inteligenței artificiale și în special în domeniul învățării automate. De asemenea am auzit cu toții predicții utopice sau apocaliptice despre ce presupun aceste progrese. Au preconizat fie nemurirea, fie sfârșitul lumii, și s-a scris mult despre ambele posibilități. Dar cele mai sofisticate inteligențe artificiale sunt încă departe de a putea soluționa problemele pe care copiii de patru ani le rezolvă cu ușurință. În ciuda numelui impresionant, inteligența artificială constă în principiu din tehnici de detectare a tiparelor statistice în seturi mari de date. Învățarea umană înseamnă mult mai mult.

Cum putem ști atât de multe despre lumea din jurul nostru? Învățăm foarte mult chiar și atunci când suntem bebeluși; copiii de patru ani știu deja lucruri despre plante, animale și mașini; dorințe, credințe și emoții; chiar și despre dinozauri și nave spațiale.

Știința ne-a prelungit cunoștințele despre lume până la ceea ce este inimaginabil de mare și ceea ce este infinitezimal de mic, până la marginea universului și începutul timpului. Și folosim aceste cunoștințe pentru a face clasificări și predicții noi, pentru a imagina posibilități noi și pentru a face să se întâmple lucruri noi în lume. Dar tot ce ajunge la noi din întreaga lume este un flux de fotoni care ne lovește retina și unde de presiune în aer care ajung la timpanele noastre. Cum învățăm atât de mult despre lume când dovezile pe care le avem sunt atât de limitate? Și cum facem toate acestea cu cele câteva kilograme de mâzgă cenușie care plutește în spatele ochilor noștri?

Cel mai bun răspuns de până acum este că creierele noastre efectuează calcule pe datele concrete, particulare și dezordonate care ajung la simțurile noastre, iar aceste calcule construiesc niște reprezentări precise ale lumii. Reprezentările par să fie structurate, abstracte și ierarhice; includ percepția unor obiecte tridimensionale, gramaticile care stau la baza limbajului și capacități mentale precum „teoria minții", care ne permit să înțelegem ce cred oamenii. Aceste reprezentări ne permit să facem o gamă largă de predicții și să ne imaginăm multe posibilități noi într-un mod creativ specific omului.

Acest tip de învățare nu este singurul tip de inteligență, dar este unul destul de important pentru ființele umane. Și este tipul de inteligență care este specialitatea copiilor cu vârste foarte fragede. Deși copiii sunt extrem de slabi la planificarea și luarea deciziilor, ei sunt creaturile care învață cel mai bine în întregul univers. O mare parte a procesului de a transforma datele în teorii se întâmplă înainte de vârsta de cinci ani.

De la Aristotel și Platon au existat două modalități de bază de a aborda problema cum știm ceea ce știm și acestea sunt încă principalele abordări în învățarea automată. Aristotel a abordat problema de jos în sus: începe cu simțurile – fluxul de fotoni și de vibrații în aer (sau pixelii ori sunetele unei imagini sau înregistrări digitale) – și vezi dacă poți extrage tipare de acolo. Această abordare a fost dusă mai departe de asociaționiști clasici, precum filosofii David Hume și J.S. Mill, și, mai târziu, de psihologii behavioriști, ca Pavlov și B.F. Skinner. În această perspectivă, caracterul abstract și structurile ierarhice ale reprezentărilor sunt aproape iluzorii sau cel puțin un epifenomen. Toată munca poate fi făcută prin asociere și detectarea tiparelor – în special dacă există suficiente date.

În timp, a existat o luptă între această abordare de jos în sus a misterului învățării și alternativa lui Platon, de sus în jos. Poate că obținem cunoștințe abstracte din date concrete pentru că știm deja multe și, în special, pentru că avem deja o multitudine de concepte abstracte de bază, mulțumită evoluției. Asemenea

cercetătorilor, putem folosi acele concepte pentru a formula ipoteze despre lume. Apoi, în loc să încercăm să extragem tipare din datele brute, putem face predicții despre cum ar trebui să arate datele dacă aceste ipoteze sunt corecte. Alături de Platon, au folosit această abordare și filosofi, și psihologi „raționaliști", precum Descartes și Noam Chomsky.

Iată exemplul de zi cu zi care ilustrează diferența dintre cele două metode: soluționarea potopului de poștă electronică nedorită. Datele constau într-o listă lungă și nesortată de mesaje din inboxul tău. Realitatea este că unele dintre aceste mesaje sunt relevante, și nu spam. Cum putem utiliza datele pentru a face diferența între ele?

Să luăm tehnica de jos în sus mai întâi. Observi că mesajele spam tind să aibă anumite caracteristici: o listă lungă de adrese, origini în Nigeria, referințe la premii de milioane de dolari sau Viagra. Problema este că și unele mesaje foarte utile pot avea aceste caracteristici. Dacă analizezi suficiente exemple de e-mailuri spam și nonspam, poți vedea nu doar că mesajele spam au acele caracteristici, dar și că ele tind să apară împreună (Nigeria plus un milion de dolari înseamnă probleme). De fapt, ar putea exista corelații subtile de nivel superior care diferențiază mesajele spam de cele utile – un anumit tipar de greșeli de scriere și adrese IP, de exemplu. Dacă detectezi aceste tipare, poți filtra mesajele spam.

Tehnicile de învățare automată de jos în sus fac acest lucru. „Botul" care învață primește milioane de exemple, fiecare cu un anumit set de caracteristici și fiecare etichetat ca fiind spam (sau altă categorie) sau nu. Computerul poate extrage tiparul de caracteristici care le diferențiază, chiar dacă este foarte subtil.

Dar dacă folosim abordarea de sus în jos? Primesc un e-mail de la editorul revistei *Journal of Clinical Biology*. Se referă la una dintre lucrările mele și spune că vrea să publice un articol scris de mine. Fără Nigeria, fără Viagra, fără milionul de dolari; e-mailul nu are niciuna dintre caracteristicile mesajelor spam. Dar, folosind ce știu deja și gândind abstract despre procesul care

produce mesajele spam, pot să-mi dau seama că acest e-mail este suspect:

1. Știu că cei care produc spam încearcă să fure bani de la alții exploatând lăcomia umană.
2. Mai știu și că revistele legitime „cu acces deschis" au început să-și acopere costurile taxându-i pe autori, și nu pe abonați, și că nu sunt specialist în biologie clinică.

Dacă alătur aceste informații, pot să produc o ipoteză nouă despre originea acelui e-mail. Este conceput pentru a-i păcăli pe autori să plătească pentru „a publica" un articol într-un jurnal fals. E-mailul a fost rezultatul aceluiași proces dubios pe care l-au suportat și alte e-mailuri spam, chiar dacă nu arată deloc ca un spam obișnuit. Pot să trag această concluzie dintr-un singur exemplu și pot continua să-mi testez ipoteza și mai departe, dincolo de orice informație din corpul e-mailului, căutând pe internet numele „editorului".

În termeni IT, am început cu un „model generativ" care include concepte abstracte precum lăcomia și înșelăciunea și descrie procesul care produce păcăleli pe e-mail. Acest lucru îmi permite să recunosc e-mailul spam clasic cu prințul nigerian, dar îmi permite, de asemenea, să-mi imaginez multe tipuri diferite de spamuri posibile. Când primesc e-mailul de la așa-zisa revistă, pot să merg înapoi pe firul raționamentului: „Acesta pare tipul de e-mail care iese dintr-un proces de generare a spamurilor".

Noul interes sporit pentru inteligența artificială vine datorită faptului că cercetătorii din domeniu au produs recent versiuni puternice și eficiente ale ambelor acestor metode de învățare. Dar nu este nimic cu adevărat nou legat de metodele însele.

Învățarea profundă ascendentă

În anii 1980, informaticienii au conceput o metodă ingenioasă de a face computerele să detecteze tipare în date: arhitectura

conecționistă sau rețeaua neuronală (partea „neuronală" era și încă este metaforică). Abordarea nu a mai fost folosită în anii 1990, dar a fost reînviată recent cu metode puternice de „învățare profundă", cum ar fi Google DeepMind.

De exemplu, îi poți da unui program de învățare profundă mai multe imagini de pe internet cu eticheta „pisică", altele cu eticheta „casă" și așa mai departe. Programul poate detecta tiparele care diferențiază cele două seturi de imagini și folosi apoi acele informații pentru a eticheta corect imagini noi. Anumite tipuri de învățare automată, denumite învățarea nesupravegheată, pot detecta tipare în date fără niciun fel de etichetă; pur și simplu, caută grupuri de caracteristici – ceea ce oamenii de știință numesc analiză a factorilor. În mașinile de învățare profundă, aceste procese sunt repetate la niveluri diferite. Unele programe pot chiar să descopere caracteristici relevante din datele brute ale pixelilor sau sunetelor; computerul poate începe prin detectarea tiparelor din imaginea brută care corespund marginilor și liniilor, apoi pot găsi tipare în acele tipare care corespund fețelor și așa mai departe.

Altă tehnică ascendentă cu istorie lungă este învățarea cu întărire. În anii 1950, B.F. Skinner, bazându-se pe lucrările lui John Watson, a intrat în istorie pentru programarea unor porumbei să facă acțiuni elaborate, chiar și să ghideze torpile aeriene către ținte (un ecou îngrijorător al inteligenței artificiale recente), oferindu-le un anumit program de recompense și pedepse. Ideea principală era că acțiunile care sunt recompensate vor fi repetate, iar cele pedepsite nu, până la obținerea comportamentului dorit. Chiar și în zilele lui Skinner, acest proces simplu, repetat de mai multe ori, putea conduce la comportamente complexe. Computerele sunt concepute să efectueze operații simple în mod repetat, la o scară ce lasă în urmă imaginația omenească, iar sistemele de calcul pot învăța competențe remarcabil de complexe în acest fel.

De exemplu, cercetătorii de la Google DeepMind au folosit o combinație de învățare profundă și învățare cu întărire pentru a

învăța un computer să joace jocuri video Atari. Computerul nu știa nimic despre cum funcționează jocurile. A început prin a acționa aleatoriu și singurele informații pe care le primea erau cum arată ecranul la un moment dat și care este scorul. Învățarea profundă a ajutat la interpretarea caracteristicilor de pe ecran, iar învățarea prin întărire recompensa sistemul pentru scoruri mari. Computerul a devenit foarte bun la câteva dintre jocurile jucate, dar, de asemenea, nu a putut să învețe nimic la altele pe care oamenii reușeau să le joace la fel de ușor ca pe celelalte.

O combinație similară de învățare profundă și învățare cu întărire a permis succesul pe care l-a obținut DeepMind cu AlphaZero, un program care a reușit să-i învingă pe jucătorii umani atât la șah, cât și la Go, echipat doar cu o cunoaștere a regulilor jocului și câteva capacități de planificare. AlphaZero mai are o caracteristică interesantă: funcționează jucând sute de milioane de jocuri cu el însuși. Pe măsură ce face acest lucru, izolează greșelile care au condus la învingeri și repetă și elaborează strategiile care au condus la victorii. Astfel de sisteme, ca și altele care implică tehnici numite rețele adversariale generative, generează date, dar și observă date.

Când ai puterea computațională pentru a aplica aceste tehnici la seturi de date foarte mari sau pe milioane de mesaje e-mail, imagini de pe Instagram sau înregistrări vocale, poți rezolva probleme care păreau foarte dificile înainte. Aceasta este sursa unei mari părți din entuziasmul care plutește acum în domeniul informaticii. Dar merită să ne amintim că acele probleme – cum ar fi recunoașterea pisicii dintr-o imagine sau că un cuvânt este pronunțat de Siri – sunt triviale pentru un bebeluș uman. Una dintre cele mai interesante descoperiri din informatică este că problemele care sunt ușoare pentru noi (cum ar fi identificarea pisicilor) sunt dificile pentru computere, mult mai dificile decât a juca șah sau Go. Computerele au nevoie de milioane de exemple pentru a clasifica obiecte pe care noi le clasificăm cu doar câteva exemple. Sistemele ascendente pot generaliza până la exemple noi, pot eticheta o imagine nouă ca fiind o pisică destul de precis,

în general. Dar fac acest lucru destul de diferit de modul în care îi generalizează pe oameni. Unele imagini aproape identice unei pisici nu vor fi identificate de noi ca pisici, iar altele care par a fi o pată aleatorie vor fi.

Modelele bayesiene descendente

Abordarea descendentă a jucat un rol important la începuturile inteligenței artificiale și, în anii 2000, a experimentat și ea o renaștere, sub forma modelelor generative probabilistice sau bayesiene.

Încercările timpurii de a utiliza această abordare au făcut față la două tipuri de probleme. În primul rând, majoritatea tiparelor de dovezi pot fi explicate în principiu prin mai multe ipoteze: este posibil ca e-mailul meu referitor la jurnal să fie veritabil, doar că nu pare. În al doilea rând, care este originea conceptelor pe care le utilizează modelul generativ? Platon și Chomsky spun că ne naștem cu ele. Dar cum putem explica modul în care învățăm ultimele concepte din știință? Sau cum înțeleg copiii tineri conceptele de dinozaur sau navă spațială?

Modelele bayesiene combină testarea modelelor generative și ipotezelor cu teoria probabilităților și abordează aceste două probleme. Un model bayesian îți permite să calculezi cât de probabil este ca o anumită ipoteză să fie adevărată, date fiind anumite informații. Și, făcând ajustări mici, dar sistematice la modelele pe care le avem deja și testându-le cu datele disponibile, putem face uneori concepte și modele noi din cele vechi. Dar aceste avantaje sunt umbrite de alte probleme. Tehnicile bayesiene te pot ajuta să alegi care ipoteză dín două este mai probabilă, și niciun sistem nu le poate lua în considerare în mod eficient pe toate. Cum decizi ce ipoteze merită testate în primul rând?

Brenden Lake de la NYU și colegii săi au utilizat aceste metode descendente pentru a rezolva altă problemă care este ușoară pentru oameni, dar extrem de dificilă pentru computere: recunoașterea caracterelor neclare scrise de mână. Privește un caracter scris pe un pergament japonez. Chiar dacă nu l-ai mai văzut niciodată, probabil

poți să-ți dai seama dacă este similar sau diferit de un caracter de pe alt pergament japonez. Probabil poți să-l desenezi și poți chiar să faci un caracter japonez fals în baza a ceea ce vezi, unul care va arăta destul de diferit de un caracter coreean sau rusesc.*

Metoda ascendentă pentru recunoașterea caracterelor scrise de mână este să dai computerului mii de exemple pentru fiecare și să-l lași pe el să extragă caracteristicile principale. În schimb, Lake și colaboratorii săi au dat programului un model general pentru cum poți desena un caracter: o mișcare din condei merge în stânga sau în dreapta; după ce termini una, începi alta și așa mai departe. Când programul vedea un anumit caracter, putea deduce cea mai probabilă secvență de mișcări care a condus la el – la fel cum eu am dedus că procesul de spam a condus la e-mailul meu dubios. Apoi poate judeca dacă un caracter nou putea rezulta din acea secvență sau din alta și putea produce el însuși un set similar de mișcări. Programul funcționa mult mai bine decât un program de învățare profundă aplicat pe exact aceleași date și oglindea îndeaproape performanța ființelor umane.

Aceste două abordări ale învățării automate au puncte forte și puncte slabe complementare. În abordarea ascendentă, programul nu are nevoie de prea multe cunoștințe la început, dar are nevoie de foarte multe date și poate generaliza doar într-un mod limitat. În abordarea descendentă, programul poate învăța din numai câteva exemple și poate face generalizări mult mai largi și mai variate, utilizând învățarea profundă pentru a implementa deducțiile bayesiene.

Succesul recent al inteligenței artificiale este în parte rezultatul extinderilor acestor idei vechi. Dar are mai mult de-a face cu faptul că, mulțumită internetului, avem mult mai multe date și, mulțumită legii lui Moore, avem mult mai multă putere de calcul pe care o putem aplica acestor date. În plus, subapreciem faptul

* Brenden M. Lake, Ruslan Salakhutdinov și Joshua B. Tenenbaum, „Human-Level Concept Learning Through Probabilistic Program Induction", revista *Science* 350, nr. 6266 (2015): 1332–38 (n.a.).

că datele pe care le avem au fost deja sortate și prelucrate de ființe umane. Imaginile cu pisici postate pe web sunt imagini canonice cu pisici – imagini pe care oamenii le-au ales deja ca fiind imagini „bune". Google Translate funcționează pentru că profită de milioane de traduceri umane și le generalizează într-o bucată nouă de text, în loc să înțeleagă cu adevărat frazele în sine.

Dar lucrul cu adevărat remarcabil despre copiii umani este că reușesc cumva să combine cele mai bune caracteristici ale fiecărei abordări, apoi trec mult mai departe de ele. În ultimii cincisprezece ani, adepții psihologiei dezvoltării au explorat modul în care copiii află structuri din date. Copiii de patru ani pot învăța din unul sau două exemple, la fel cum învață un sistem descendent, și pot generaliza concepte foarte diferite. Dar pot învăța și concepte și modele noi de la datele în sine, așa cum face un sistem ascendent.

De exemplu, în laboratorul nostru, dăm copiilor tineri un „detector de blicheți"* – o mașină nouă ce trebuie descoperită, una pe care nu au mai văzut-o niciodată. Este o cutie care se aprinde și redă muzică când pui anumite obiecte în ea, și nu altele. Le dăm copiilor unul sau două exemple pentru cum funcționează mașina, arătându-le că, de exemplu, două cuburi roșii o activează, în timp ce o combinație între verde și galben nu o activează. Chiar și copiii în vârstă de optsprezece luni pricep imediat principiul general, și anume că cele două obiecte trebuie să fie identice pentru a activa cutia, și generalizează acel principiu la exemple noi: de exemplu, vor alege două obiecte care au aceeași *formă* pentru a face mașina să funcționeze. În alte experimente, am arătat despre copii că pot pricepe că anumite proprietăți ascunse invizibile fac mașina să funcționeze sau că mașina funcționează în baza unui principiu logic abstract."

---

* „Blicket" este un concept introdus de Nancy Soja ce presupune un tip de obiect nou, cu anumite proprietăți ce pot fi descoperite de om prin anumite experimente legate de cauzalitate sau percepție (n.t.).

** A. Gopnik, T. Griffiths și C. Lucas, „When Younger Learners Can Be Better (or at Least More Open-Minded) Than Older Ones", *Current Directions in Psychological Science* 24, nr. 2 (2015): 87–92 (n.a.).

Putem verifica acest principiu și în procesul de învățare obișnuit al copiilor. Copiii tineri învață rapid teorii intuitive abstracte din biologie, fizică și psihologie în aproape același fel în care o fac și oamenii de știință adulți, chiar și cu relativ puține date.

Realizările remarcabile ale învățării automate ale sistemelor recente de inteligență artificială – atât ascendente, cât și descendente – au loc într-un spațiu restrâns și bine definit de ipoteze și concepte – un set precis de componente de joc și mutări, un set predeterminat de imagini. Comparativ, copiii și oamenii de știință își schimbă uneori radical conceptele, realizând schimbări de paradigmă în loc să ajusteze pur și simplu conceptele pe care le au deja.

Copiii de patru ani pot recunoaște imediat pisicile și pot înțelege cuvintele, dar pot, de asemenea, face deducții creative și surprinzătoare care trec mult dincolo de experiențele lor. Propriul meu nepot mi-a explicat recent, de exemplu, că, dacă un adult vrea să devină din nou un copil, ar trebui să încerce să nu mănânce legume sănătoase, deoarece legumele sănătoase îi fac pe copii să crească și să se transforme în adulți. Acest tip de ipoteză, una plauzibilă la care nu s-ar gândi un adult, este caracteristic copiilor tineri. De fapt, eu și colegii mei am arătat sistematic că preșcolarii sunt mai buni la crearea de ipoteze improbabile decât copiii mai în vârstă sau decât adulții.* Nu avem aproape nicio idee cum este posibil acest tip de învățare creativă și inovare.

Privind ce fac copiii însă, programatorii pot face observații utile despre direcțiile învățării pentru computere. Două caracteristici ale procesului de învățare al copiilor sunt izbitoare. Copiii învață activ, nu absorb pasiv date, așa cum face inteligența artificială. La fel cum experimentează oamenii de știință, copiii sunt motivați intrinsec să extragă informații din lumea înconjurătoare prin explorarea și joaca lor fără de sfârșit. Studiile recente arată că

* A. Gopnik și colaboratorii, „Changes in Cognitive Flexibility and Hypothesis Search Across Human Life History from Childhood to Adolescence to Adulthood", *PNAS* 114, nr. 30 (2017): 7892–99 (n.a.).

această explorare este mai sistematică decât pare și este bine adaptată pentru a găsi dovezi convingătoare pentru a suporta formarea ipotezelor și alegerea teoriilor.* Integrarea curiozității în mașini și permisiunea de a interacționa activ cu lumea pot fi un drum mai realist către învățarea diversificată.

În al doilea rând, copiii, spre deosebire de inteligențele artificiale existente, învață social și cultural. Oamenii nu învață în izolare, ci se folosesc de înțelepciunea acumulată a generațiilor trecute. Studiile recente arată că până și preșcolarii învață prin imitație și ascultând mărturiile altora. Dar nu își ascultă pur și simplu pasiv profesorii. În schimb, interiorizează informații de la alte persoane într-un mod remarcabil de subtil și sensibil, făcând deducții complexe despre locul din care vine informația și cât este de fiabilă, integrându-și sistematic propriile experiențe în ceea ce aud.**

„Inteligența artificială" și „învățarea automată" sună înspăimântător. Și, în anumite feluri, sunt. Aceste sisteme sunt utilizate pentru a controla arme, de exemplu, și de asta chiar ar trebui să ne temem. Totuși, prostia naturală poate cauza dezastre mult mai mari decât inteligența artificială; noi, oamenii, va trebui să fim mult mai inteligenți decât am fost în trecut pentru a gestiona corespunzător noile tehnologii. Dar nu există baze solide nici pentru viziunea apocaliptică, nici pentru cea utopică, în care inteligența artificială îi înlocuiește pe oameni. Cât timp nu vom rezolva paradoxul elementar al învățării, cele mai bune inteligențe artificiale nu vor putea concura cu un copil obișnuit de patru ani.

---

* L. Schulz, „The Origins of Inquiry: Inductive Inference and Exploration in Early Childhood", *Trends in Cognitive Sciences* 16, nr. 7 (2012): 382–89 (n.a.).

** A. Gopnik, *The Gardener and the Carpenter* (New York: Farrar, Straus & Giroux, 2016), capitolele 4 și 5 (n.a.)

Capitolul 22
Algoritmicienii visează la obiectivitate

Peter Galison

Peter Galison este istoric al științei; profesor universitar Joseph Pellegrino și cofondator al inițiativei Black Hole de la Universitatea Harvard; și autorul cărții *Einstein's Clocks Poincaré's Maps: Empires of Time.*

În calitate de istoric al științei, Peter Galison se concentrează, în general, pe intersecția teoriei cu experimentul.

„De ani buni, îmi ghidez activitatea după confruntarea ciudată a ideilor abstracte cu obiecte extrem de concrete", mi-a spus el cândva, explicând care este părerea lui despre ceea ce face. La întâlnirea din Washington, Connecticut, a discutat despre tensiunile din Războiul Rece dintre ingineri (ca Wiener) și administratorii proiectului Manhattan (ca Oppenheimer): „Când Wiener ne avertizează despre pericolele ciberneticii, pe de o parte, încearcă să concureze cu limbajul neobișnuit pe care îl folosesc oamenii ca Oppenheimer: «Când am văzut explozia de la Trinity, m-am gândit la *Bhagavad Gita* – Eu sunt moartea, distrugătoarea lumilor». Acest sens, și anume că fizica poate vorbi despre natura universului și politicile militare aeriene,

era respingător și seducător. Într-un fel, se poate vedea acest lucru în mod repetat în ultimele decenii – nanoștiințele, ADN recombinant, cibernetica: «Vă raportez despre știința care aduce cu ea promisiunea salvării și pericolul anihilării – și trebuie să fiți atenți, pentru că vă poate omorî». Este o poveste foarte seducătoare și este repetată în inteligența artificială și în robotică".

Când aveam douăzeci și patru de ani și am intrat prima dată în contact cu ideile lui Wiener, apoi i-am cunoscut colegii la întâlnirile de la MIT pe care le descriu în introducerea cărții, eram foarte puțin interesat de avertismentele sau admonestările lui Wiener. Curiozitatea mea era împinsă de la spate de natura puternică și radicală a perspectivei sale asupra vieții, bazate pe teoria matematică a comunicațiilor în care mesajul este neliniar: după Wiener, „conceptele noi ale comunicării și controlului au implicat o nouă interpretare a omului, a cunoștințelor omului despre univers și a societății". Și acest lucru a condus la prima mea carte, care a luat teoria informației – teoria matematică a comunicării – ca model pentru toată experiența umană.

Într-o conversație recentă, Peter mi-a spus că începe să scrie o carte – despre construire, prăbușire și gândire – care analizează natura de cutie neagră a ciberneticii și cum reprezintă aceasta ceea ce percepe ca „transformarea fundamentală a învățării, învățării automate, ciberneticii și sinelui".

În a doua cea mai bună carte a sa, marele matematician medieval al-Khwarizmi a descris noua formă de aritmetică de origine indiană. Numele său, în scurt timp corelat cu „algorismus" (în latina medievală târzie), a ajuns să desemneze procedurile aplicate numerelor, alunecând treptat către „algoritm" (pe modelul „logaritm"), în franceză și apoi în engleză. Dar îmi place ideea de algoritmician modern, chiar dacă această idee nu-i place și corectorului meu gramatical. Prin acest termen înțeleg o persoană profund suspicioasă cu privire la intervenția judecății umane, cineva care folosește această judecată pentru a încălca normele fundamentale a ceea ce înseamnă să fii obiectiv (și, prin urmare, științific).

Aproape de sfârșitul secolului XX, o lucrare scrisă de doi psihologi de la Universitatea din Minnesota a rezumat o literatură vastă care a tulburat mult timp apele prognozelor. O parte, spuneau ei, a ținut prea mult timp cu dinții – și, în cele din urmă, fără etică – de „metoda clinică" de prognoză, care punea accent pe tot ce este subiectiv: „informal", „în cap" și „impresionist". Acești clinicieni erau oameni (spuneau psihologii) care credeau că pot să-și studieze subiecții cu o grijă meticuloasă, se pot aduna în comisii și pot face predicții pe bază de judecată cu privire la recidivismul criminal, succesul la facultate, rezultatele medicale și altele asemenea. Cealaltă parte, au continuat psihologii, întrupa tot ce nu întrupau clinicienii, îmbrățișând obiectivul: „formal", „mecanic", „algoritmic". Autorii au considerat că acest lucru stă la rădăcina întregului triumf al științei de după Galilei. Nu doar că

știința beneficia de pe urma statisticii, dar și, în mare măsură, știința era statisticianul mecanic. Răsfoind 136 de studii pe tema predicțiilor, din domenii diverse de la sentințele judecătorești la psihiatrie, autorii au arătat că, în 128 dintre ele, predicțiile ce conțineau tabele statistice, o ecuație de regresie multiplă sau o judecată algoritmică au egalat sau depășit în acuratețe pe cele care au folosit abordarea subiectivă.

Au continuat prin catalogarea a șaptesprezece justificări eronate pentru susținerea abordării clinice. Aceștia erau trântorii care-și vedeau propriile interese și care se temeau că mașinile îi vor lăsa fără locurile lor comode de muncă. Alții nu aveau educația necesară să urmărească argumentele statistice. Un grup nu avea încredere în formalizarea matematicii; alții au criticat sever ceea ce ei considerau că este o „dezumanizare statistică"; iar alții au spus că scopul este înțelegerea, nu predicția. Dar, oricare ar fi motivațiile, rezumatul a conchis că este de-a dreptul imoral să refuzi puterea obiectivității asupra subiectivității, a algoritmilor asupra judecății experților.*

Perspectiva algoritmicianului a câștigat teren. Anne Milgram a fost procuror general al statului New Jersey din 2007 până în 2010. Când a preluat postul, a vrut să știe pe cine aresta, acuza și întemnița statul și pentru ce infracțiuni. La momentul respectiv, raportează ea într-o conferință TED ulterioară, datele sau statisticile erau aproape inexistente. Impunând predicțiile statistice, continuă ea, forțele de ordine din Camden în timpul mandatului său au reușit să reducă rata crimelor comise cu 41%, salvând 37 de vieți, reducând rata totală a crimelor la 26 de procente. După ce s-a alăturat Fundației Arnold în calitate de vicepreședinte pentru justiție penală, a stabilit o echipă de administratori de date și statisticieni pentru a crea un instrument de evaluare a riscurilor;

* William M. Grove and Paul E. Meehl, „Comparative Efficiency of Informal (Subjective, Impressionistic) and Formal (Mechanical, Algorithmic) Prediction Procedures: The Clinical- Statistical Controversy", *Psychology, Public Policy, and Law* 2, no. 2 (1996): 293– 323 (n.a.).

de fapt, a interpretat misiunea echipei ca fiind aceea de a decide cum vor fi întemnițate „persoanele periculoase" și cum vor fi eliberate cele nepericuloase. „Motivul care stătea în spatele acestei decizii", a continuat Milgram, „este modul în care luăm deciziile. Judecătorii au cele mai bune intenții când iau aceste decizii cu privire la riscuri, dar le iau subiectiv. Erau ca evaluatorii de talente din baseball de acum douăzeci de ani care încercau să își folosească instinctul și experiența pentru a încerca să stabilească ce riscuri prezintă cineva. Erau subiectivi și știm ce se întâmplă când oamenii iau decizii subiective, anume că ne înșelăm adesea". Echipa sa a stabilit peste nouă sute de factori de risc, dintre care nouă erau niște predictori foarte buni. Cele mai urgente întrebări pentru echipă erau: o persoană va comite o nouă infracțiune? Persoana respectivă va comite un act de violență? Cineva va reveni în fața instanței? Avem nevoie, a concluzionat Milgram, de o „măsură obiectivă pentru risc" care ar trebui aplicată de judecata judecătorilor. Știm că procesul statistic algoritmic funcționează. Acesta este motivul pentru care „Google este Google", spune ea, și motivul pentru care vânătoarea de talente câștigă meciurile.*

Algoritmicienii au câștigat. Ne-am obișnuit cu ideea că protocoalele și datele pot și trebuie să ne ghideze în acțiunile noastre de zi cu zi, de la mementouri despre locurile în care trebuie să ajungem până la probabilitatea de a fi victima unei infracțiuni. Până acum, conform literaturii, dimensiunile legală, etică, oficială și economică ale algoritmilor sunt toate cvasiinfinite. Aș vrea să mă concentrez mai mult pe un cântec de sirenă al algoritmilor: promisiunea lor de obiectivitate.

Obiectivitatea științifică are o istorie. Acest lucru poate părea surprinzător. Nu este corectă noțiunea exprimată mai sus de psihologii din Minnesota? Nu este obiectivitatea coextensivă cu știința însăși? Aici merită să facem un pas înapoi pentru a reflecta asupra tuturor virtuților epistemice cărora le-am putea acorda

* TED Talk, January 2014, https:// www.ted.com/ speakers/ anne_ milgram (n.a.).

valoare în activitatea științifică. Calificarea pare ceva bun de avut; la fel și predicția, explicația, unificarea, precizia, acuratețea, siguranța și utilitatea pedagogică. În cea mai bună dintre lumile posibile, aceste virtuți epistemice ar trage toate în aceeași direcție. Dar nu fac asta – nu mai mult decât coincid în mod necesar virtuțile noastre etice. Recompensarea oamenilor în funcție de nevoile lor poate intra în conflict cu recompensarea oamenilor în funcție de competențe. Egalitate, echitate și meritocrație – etica, într-un anumit sens, este legată de adjudecarea bunurilor conflictuale. Prea des uităm că acest conflict există și în știință. Proiectează un instrument pentru a fi cât se poate de sensibil și adesea acesta va fluctua mult, făcând imposibilă repetarea unei măsurări.

„Obiectivitatea științifică" a pătruns atât în practica, precum și în nomenclatura științifică după prima treime a secolului al XIX-lea. Se poate vedea clar în atlasele științifice care au pus la baza oamenilor de știință obiectele de bază ale specialității lor: au existat și există atlase ale mâinii, ale craniului, ale norilor, cristale, flori, tablouri de cameră, emulsii nucleare și boli ale ochilor. În secolul al XVIII-lea, era evident că nu vom înfățișa o anumită frunză de trifoi, arsă de soare și mestecată de omizi, pe care o găsim în fața casei, într-un atlas. Nu, ținta era – dacă erai un filosof natural genial precum Goethe, Albinius sau Cheselden – să observăm natura, dar apoi să perfecționăm obiectul observat, să-l abstractizăm vizual până la ideal. Să luăm, de exemplu, un schelet, să-l vizualizăm printr-o cameră lucidă și să-l desenăm cu atenție, apoi să corectăm „imperfecțiunile". Avantajul acestei despărțiri a cortinelor de simplă experiență era clar: punea la dispoziție un ghid universal, unul neatașat de inconsecvențele unei variații individuale.

Pe măsură ce volumul științei a crescut, și numărul oamenilor de știință la fel, partea proastă a idealizării a devenit mai clară. Era ceva când Goethe înfățișa „planta primordială" sau „insecta primordială". Dar era cu totul altceva să avem miriade de cercetători diferiți care-și fixează imaginile în moduri diferite și uneori contradictorii. Gradual, începând cu anii 1830, începem să

observăm ceva nou: pretenția că realizarea imaginii era făcută cu o intervenție umană minimă, că erau respectate protocoale. Acest lucru putea însemna conturarea unei frunze cu creionul sau apăsarea acesteia în tuș și transferarea ei pe pagină. Mai însemna și că ne mândream cu înfățișarea imaginii prin microscop a unui obiect natural chiar și cu imperfecțiunile sale. Aceasta era o idee radicală: fulgi de zăpadă înfățișați fără o simetrie hexagonală perfectă, distorsiunea culorii pe marginea unei lentile de microscop, țesut rupt în jurul marginilor în procesul pregătirii sale.

Obiectivitatea științifică a ajuns să însemne că reprezentările noastre erau executate prin eliminarea intervenției – chiar dacă asta însemna să reproducem culoarea galbenă de lângă marginea imaginii sub microscop, în ciuda faptului că omul de știință știa că decolorarea era cauzată de lentilă, și nu o caracteristică a obiectului cercetat. Avantajul obiectivității era clar: înlocuia dorința de a vedea o teorie realizată sau o perspectivă acceptată confirmată. Dar obiectivitatea avea costurile sale. Pierdeam reprezentarea artistică precisă, care se putea învăța ușor, colorată și care cuprindea întreaga profunzime a câmpului în cazul unui cadavru disecat. Obțineam o fotografie încețoșată, cu o profunzime slabă a câmpului, alb-negru, pe care niciun student la medicină (și nici mulți colegi din medicină) nu o puteau folosi pentru a învăța și compara cazuri. Cu toate acestea, o lungă parte din secolul al XIX-lea, virtutea obiectivității cu intervenție redusă era în creștere.

Începând cu anii 1930, obiectivitatea științifică strictă din reprezentarea științifică a început să aibă probleme. În catalogarea spectrelor stelare, de exemplu, niciun algoritm nu putea concura cu observatorii antrenați, care le puteau sorta cu o acuratețe și replicabilitate mult mai mari decât orice procedură care urmează reguli. La sfârșitul anilor 1940, doctorii au început să învețe să citească electroencefalogramele. Era necesară o părere de expert pentru sortarea diverselor tipuri de citiri ale atacurilor cerebrale, în timp ce niciuna dintre încercările dinainte de a utiliza analiza frecvențelor nu se potrivea cu judecata lor. Magnetogramele solare – cartografierea câmpurilor magnetice de pe Soare – necesita

din partea expertului să diferențieze semnalul real de artefactele care luau naștere din cauza instrumentelor de măsurare. Chiar și fizicienii care se ocupă de particule au recunoscut că nu puteau programa un computer pentru a sorta anumite tipuri de urme în categoriile lor speciale; era necesară o părere de profesionist.

Nu trebuie să existe confuzie aici: aceasta nu era o întoarcere la geniul invocat al unui idealist din secolul al XVIII-lea. Nimeni nu credea că te puteai antrena să devii un Goethe care singur printre alți cercetători putea alege forma universală și ideală a unei plante, a unei insecte sau a unui nor. Expertiza se poate deprinde – poți face un curs pentru a învăța să faci judecăți de expert în ceea ce privește electroencefalogramele, spectrele stelare sau urmele din camerele cu bule; din păcate, nimeni nu s-a gândit că ai putea absolvi un curs care te-ar putea face un observator excepțional. Nu există niciun drum regal către a deveni un Goethe. În nenumărate atlase științifice se pot vedea argumente explicite că factorii „subiectivi" trebuie să facă parte din activitatea științifică necesară pentru a crea, clasifica și interpreta imaginile științifice.

Ce vedem în atât de multe pretenții ale algoritmicienilor este o dorință uriașă de a găsi obiectivitatea științifică fix prin abandonarea judecății și bizuirea pe proceduri mecanice – în numele obiectivității științifice. Multe state americane au legiferat utilizarea unor algoritmi de stabilire a sentințelor și cauțiunilor. Mai bine o mașină, se argumentează, decât inconsecvențele judecății unui judecător.

Iată deci o avertizare din partea științelor. Proceduralismul algoritmic fără intervenție umană a avut într-adevăr perioada de înflorire în secolul al XIX-lea și, bineînțeles, încă joacă un rol în multe dintre încercările tehnice și științifice de succes. Dar ideea că obiectivitatea mecanică, interpretată drept control strict de sine, urmează o curbă simplă și monotonă, crescând de la clinicianul impresionist și slab la o statistică externalizată bună, nu oferă răspuns la istoria mai interesantă și nuanțată a științelor.

Există o lecție și mai importantă pe care o învățăm de la științe. Obiectivitatea mecanică este o virtute științifică printre

altele, iar științele naturale învață această lecție destul de des. Trebuie să facem la fel și în domeniile științelor legale și sociale. Ce se întâmplă, de exemplu, când algoritmul secret și brevetat trimite o persoană la închisoare timp de zece ani și alta timp de cinci ani pentru aceeași infracțiune? Rebecca Wexler, profesor consultant la proiectul societății informatice de la Facultatea de Drept Yale, a explorat această întrebare și costul uriaș pe care algoritmii secretelor comerciale îl impun pentru posibilitatea de apărare legală echitabilă. Într-adevăr, dintr-o varietate de motive, organele legale nu vor fi dispuse să împărtășească algoritmii utilizați pentru producția de ADN, substanțe chimice, identificarea amprentelor, ceea ce pune apărarea într-o poziție mult slăbită pentru a-și susține pledoaria. În sala de judecată, obiectivitatea, secretele comerciale și transparența legală pot trage în direcții opuse. Îmi aduce aminte de un moment din istoria fizicii. Imediat după cel de-al Doilea Război Mondial, giganții în producția de film Kodak și Ilford au perfecționat un film care putea fi utilizat pentru a descoperi interacțiunile și descompunerile particulelor elementare. Fizicienii erau încântați, bineînțeles – până când companiile de film le-au spus despre compoziția filmului că este un secret comercial, astfel încât cercetătorii nu au putut avea niciodată încredere deplină că înțeleg corect procesele pe care le studiază. Dacă dotăm lucrurile cu niște cutii negre care nu pot fi deschise, oamenii de știință vor juca un joc periculos, cu atât mai mult justiția penală.

Alți critici au evidențiat cât de periculos este să ne bizuim pe formulările unei persoane acuzate (sau condamnate) sau pe alte variabile care pot deveni cu ușurință, în interiorul cutiei negre a stabilirii sentinței pe bază de algoritmi, o împuternicire pentru rasă. Prin urmele experienței noastre de zi cu zi, ne-am obișnuit cu faptul că securitatea de la aeroport este diferită pentru copiii sub doisprezece ani și pentru adulții peste șaptezeci și cinci de ani. Ce factori dorim să aibă algoritmicienii în procedurile lor adesea ascunse? Educație? Venit? Istoricul locurilor de muncă? Ce a citit, privit, vizitat sau cumpărat o persoană? Contactele anterioare cu

legea? Cum dorim să pondereze algoritmicienii acești factori? Funcțiile analitice predictive predicate pe obiectivitatea mecanică au prețul lor. Uneori poate fi un preț ce merită plătit; alteori acel preț ar fi devastator pentru societatea echitabilă pe care dorim să o avem.

În general, pe măsură ce convergența algoritmilor și a Big Data ne controlează o parte din ce în ce mai mare a vieții, merită să ne amintim aceste două lecții din istoria științelor: judecata nu este coaja decorticată a unei obiectivități acum pure a controlului de sine. Și obiectivitatea mecanică este o virtute care concurează cu altele, și nu chintesența unei întreprinderi științifice. Acestea sunt lecții de ținut minte, chiar dacă algoritmicienii visează la obiectivitate.

Capitolul 23
Drepturile mașinilor

George M. Church

George M. Church este profesor de genetică Robert Winthrop la Facultatea de Medicină de la Harvard, profesor de științe ale sănătății și tehnologie la MIT Harvard și coautor (împreună cu Ed Regis) al cărții *Regenesis: How Synthetic Biology Will Reivent Nature and Ourselves.*

În ultimul deceniu, ingineria genetică a prins din urmă informatica cu privire la modul în care noile inițiative științifice ne modelează viețile. Inginerul genetician George M. Church, un pionier al revoluției de citire și scriere a biologiei, este un jucător important în acest peisaj nou de idei. Privește corpul ca pe un sistem de operare, inginerii luând locul biologilor tradiționali în reutilarea componentelor învechite ale organismelor (de la atomi la organe), aproape în același fel ca la sfârșitul anilor 1970, când inginerii lucrau pentru a crea primul computer personal, asamblând plăci de circuite, unități de disc, monitoare etc. George a creat și este administratorul proiectului Personal Genome, care conduce singura sursă de informații cu acces

deschis la datele genomului uman, despre mediu și trăsături (GET), revigorând industria străveche, acum în creștere, a ADN-ului.

Activitatea sa a fost foarte importantă în punerea bazelor inițiativei BRAIN (Brain Research through Advancing Innovative Neurotechnologies – Cercetarea creierului prin neurotehnologii inovatoare avansate) din 2013 a președintelui Obama – în ajutorul îmbunătățirii creierelor omenești până la punctul în care, pentru o mare parte din ceea ce avem nevoie, să nu mai avem nevoie de inteligențe artificiale (potențial periculoase). „Este posibil ca unele dintre proiectele Inițiativei BRAIN să ne permită să construim creiere omenești care se potrivesc mai bine cu etica noastră și sunt capabile să execute sarcini avansate, asemenea inteligenței artificiale", spunea George. „Cea mai sigură cale de departe este ca oamenii să reușească să facă toate sarcinile pe care ar vrea să le delege mașinilor, dar nu suntem încă stabili pe această cale supersigură."

Mai recent, utilizarea sa de pionierat de o importanță crucială a enzimei CRISPR (precum și a metodelor mai bune decât CRISPR) pentru a edita genele celulelor umane este uneori omisă de media, atunci când se spune povestea originilor CRISPR.

Atitudinea lui George la adresa viitoarelor forme ale inteligenței artificiale generale este prietenoasă, după cum se vede în eseul care urmează. În același timp, nu pierde niciodată din vedere problema siguranței inteligenței artificiale. Pe această temă, observa recent: „Principalul risc în domeniul inteligențelor artificiale, din punctul meu de vedere, nu este atât de mult dacă putem înțelege matematic ce gândesc, ci dacă suntem capabili să le învățăm comportamentul etic. Abia dacă putem să-i învățăm pe semenii noștri comportamentul etic".

În 1950, cartea lui Norbert Wiener *The Human Use of Human Beings* era un apogeu de viziune și speculație spunând că:

> Mașina, asemenea djinnului, care poate învăța și lua decizii în baza a ceea ce a învățat, nu va fi obligată în niciun fel să ia decizii pe care noi ar fi trebuit să le luăm sau care să fie acceptabile pentru noi... Indiferent dacă ne încredințăm deciziile noastre mașinilor de metal sau mașinilor din carne și oase, care sunt birouri, laboratoare vaste, armate și corporații... este târziu, iar alegerea dintre bine și rău bate la ușă.

Dar acesta era deznodământul cărții sale și ne-a lăsat în așteptare de șaizeci și opt de ani, lipsindu-i nu numai prescripțiile și proscripțiile, dar chiar și o „formulare a problemei" bine articulată. De atunci am văzut avertismente similare cu privire la amenințarea mașinilor noastre, chiar și sub forma strigătului către mase, prin filme precum *Colossus: The Forbin Project* (1970), *The Terminator* (1984), *The Matrix* (1999) și *Ex Machina* (2015). Dar acum este vremea unei actualizări majore, cu perspective noi și proaspete – notabil concentrate pe generalizări ale drepturilor „omului" și ale nevoilor noastre existențiale.

Îngrijorările au avut tendința să se concentreze pe „noi contra lor [roboților]" sau „mâzga gri [nanotehnologia]" sau

„monoculturile de clone [bio]". Pentru a extrapola tendințele curente: ce-ar fi dacă am putea face sau crește aproape orice și am putea proiecta orice nivel de siguranță și eficacitate am dori? Orice ființă gânditoare (făcută din orice aranjament de atomi) poate avea acces la orice tehnologie.

Probabil ar trebui să ne concentrăm mai puțin pe noi-contra-lor și mai mult pe drepturile ființelor simțitoare în fața emergenței unei diversități de minți fără precedent. Ar trebui să exploatăm această diversitate pentru a minimiza riscurile existențiale globale, cum ar fi supervulcanii și asteroizii.

Dar, ar trebui oare să spunem „ar trebui"? (Precizări necesare: în acest caz și în multe altele, când un tehnolog descrie o cale a societății care „se poate", „se va" sau „ar trebui să" se întâmple, acest lucru nu înseamnă neapărat că acestea sunt și preferințele autorului. Poate reflecta avertizarea, nesiguranța și/sau evaluarea detașată.) Roboticianul Gianmarco Veruggio și alții au ridicat probleme de roboetică încă din 2002; Departamentul de Comerț și Industrie din Marea Britanie și Institutul pentru Viitor, o subcompanie a RAND, au ridicat probleme legate de drepturile roboților încă din 2006.

„Este versus ar trebui"

Este de bun-simț să spui că știința se preocupă de „este", și nu de „ar trebui". Perspectiva „magisteriei nesuprapunerii" a lui Stephen Jay Gould argumentează că faptele trebuie să fie complet diferite de valori. Similar, documentul *Știință și creaționism* din 1999, din U.S. National Academy of Sciences, observa că „știința și religia ocupă două tărâmuri separate". Această împărțire a fost criticată de biologul evoluționist Richard Dawkins, de mine și de alții. Îl putem discuta pe „ar trebui" dacă este încadrat ca „ar trebui să facem X pentru a putea obține Y". Ce Y ar trebui să aibă o prioritate mai mare nu este ales neapărat prin vot democratic, dar poate fi ales prin vot darwinian. Sistemele de valori și suișurile și

coborâșurile religiei diversifică, deviază și converg la fel ca speciile vii: sub imperiul selecției. „Valoarea" finală (acel „ar trebui") este supraviețuirea genelor și a memelor.

Puține religii spun că nu există nicio conexiune între ființa noastră fizică și lumea spirituală. Miracolele sunt documentate. Conflictele dintre doctrina bisericească și Galileo sau Darwin sunt rezolvate în cele din urmă. Credința și etica se găsesc din abundență din specia noastră și pot fi studiate folosind metode științifice, printre care, dar nu exclusiv, fRMN, medicamente psihoactive, chestionare etc.

Foarte practic, avem de găsit regulile etice care trebuie construite, învățate sau alese probabilistic pentru mașini din ce în ce mai inteligente și diverse. Avem o întreagă serie de probleme ale tramvaiului. La ce număr de oameni care-și așteaptă moartea ar trebui computerul să decidă să schimbe macazul la o persoană? În cele din urmă, aceasta poate fi o problemă de învățare profundă – una în care baze de date uriașe de fapte și contingențe pot fi luate în considerare, unele aparent departe de etica discutată.

De exemplu, computerul poate deduce că persoana care scapă de moarte dacă tramvaiul este lăsat să-și urmeze cursul este un recidivist terorist condamnat, plin de patogeni care pot distruge lumea, sau un POTUS sfânt – sau parte dintr-un lanț de evenimente mult mai elaborat în realități alternative detaliate. Dacă una dintre aceste descrieri al problemei pare paradoxală sau ilogică, este posibil ca autorii Problemei Tramvaiului să fi ajustat greutățile în fiecare parte astfel încât deciziile ezitante să fie inevitabile.

Alternativ, se poate folosi redirecționarea atenției pentru a măslui sistemul, astfel încât modurile de eroare să nu fie la nivelul atenției. De exemplu, în Problema Tramvaiului, adevărata decizie etică a fost făcută cu câțiva ani înainte, când pietonii au primit acces pe șine sau chiar înainte de asta, când am votat să cheltuim mai mult pe divertisment decât pe siguranța publică. Întrebări care par la început străine și tulburătoare, cum ar fi „Cine deține noile minți și cine plătește pentru greșelile lor?", sunt similare

legilor bine stabilite cu privire la cine deține și plătește pentru greșelile unei corporații.

Pantele alunecoase

Putem (supra)simplifica etica pretinzând că anumite scenarii nu se vor întâmpla. Provocările tehnice sau liniile roșii aprinse care nu pot fi trecute sunt reconfortante, dar realitatea este că, odată ce beneficiile depășesc riscurile (chiar și puțin sau pentru scurt timp), liniile roșii se mută. Chiar înainte de nașterea lui Louise Brown, în 1978, mulți oameni își făceau griji că „va deveni un mic monstru, cumva, deformată, că va fi ceva în neregulă cu ea"*. Puțini mai văd fertilizarea in vitro așa în prezent.

Ce tehnologii lubrifiază panta către conștiința compusă? Nu este vorba doar despre algoritmii de învățare automată profundă cu Big Iron. Am reușit să facem rozătoarele să fie mult mai bune la o diversitate de sarcini cognitive, precum și la manifestarea altor trăsături relevante, cum ar fi persistența și anxietatea scăzută. Va fi adevărat, oare, și pentru animalele care se apropie de inteligența umană? Multe arată recunoașterea sinelui în testul oglinzii – cimpanzei, bonobo, urangutani, anumiți delfini și anumite balene, precum și coțofenele.

Chiar și linia roșie aprinsă pentru manipularea oamenilor de către oameni arată multe semne de mișcare sau de rupere cu totul. Peste 2 300 de studii clinice aprobate pentru terapia cu gene sunt în desfășurare în întreaga lume. Un obiectiv medical major este tratarea sau împiedicarea scăderii capacităților cognitive, în special în lumina demografiei globale care arată o tendință rapidă de îmbătrânire. Unele tratamente pentru scăderea capacităților cognitive vor include stimulatori cognitivi (medicamente, gene, celule, transplanturi, implanturi și altele). Acestea vor fi utilizate pentru alte indicații terapeutice decât cele indicate de prospecte.

* „Then, Doctors «All Anxious» About Test- tube Baby", http://edition.cnn.com/2003/HEALTH/parenting/07/25/cnna.copperman (n.a.).

Regulile competițiilor atletice (de exemplu, interzicerea augmentării cu steroizi sau eritropoietină) nu se aplică, așadar, competițiilor intelectuale din lumea reală. Toate progresele realizate în domeniul scăderii capacităților cognitive sunt în joc pentru indicațiile neaprobate.

Altă frontieră pentru folosirea oamenilor de către oameni este reprezentată de „organoidele cerebrale". Acum putem accelera biologia dezvoltării. Procese care în mod normal durau luni de zile se pot realiza acum în patru zile în laborator folosind rețetele potrivite de factori de transcripție. Putem face creiere ce recapitulează diferențele dintre oamenii născuți cu capacități cognitive aberante (de exemplu, microcefalie) cu o fidelitate din ce în ce mai mare. Vascularizarea corespunzătoare (vene, artere și capilare) ce lipsea din succesele anterioare este adăugată acum, permițând organoidelor cerebrale să depășească limita anterioară de submicrolitru, cu probabilitatea de a depăși mărimea de 1,2 litri a creierelor umane moderne (sau chiar a creierelor de elefant, de cinci litri, sau de balenă, de 8 litri).

Computere convenționale versus hibrizi bioelectrici

Pe măsură ce ne apropiem de un nou hop în miniaturizarea din Legea lui Moore (care, cu siguranță, nu este un zid de netrecut), vedem limitele stocasticii atomilor dopanți în plăci de siliciu și limitele metodelor de fabricare a grinzilor la mărimi de aproximativ zece nanometri. Problemele legate de consumul de energie sunt și ele clare: marele Watson, câștigătorul emisiunii *Jeopardy!*, a folosit 85 000 de wați în timp real, în vreme ce creierele omenești foloseau 20 de wați fiecare. Să fim cinstiți, corpul omenesc are nevoie de 100 de wați pentru a funcționa și de 20 de ani pentru a fi construit, deci aproximativ 6 trilioane de jouli de energie pentru a „produce" un creier uman matur. Costul de producție al unui sistem de calcul Watson este similar. Prin urmare, de ce nu înlocuiesc oamenii computerele?

În primul rând, creierele concurenților de la *Jeopardy!* făceau mult mai mult decât să colecteze informații – multe dintre acestea ar fi considerate doar distrageri de Watson (de exemplu, controlul cerebral al zâmbitului). Alte părți însemnau o ieșire din cutie cu o transcendență de neconceput pentru Watson, așa cum vedem în cele cinci lucrări ale lui Einstein din *annus mirabilis* 1905. De asemenea, oamenii consumă mai multă energie decât minimul (100 de wați) necesar pentru viață și reproducere. Oamenii din India consumă în medie 700 de wați de persoană; în Statele Unite, această medie este de 10 000. Ambele medii sunt sub consumul de 85 000 de wați al lui Watson. Computerele pot începe să semene cu noi prin tehnica de calcul neuromorfică, posibil de 1 000 de ori mai bine. Dar și creierele omenești pot deveni mai eficiente. Organoidul creier în sticlă se poate apropia de limita de 20 de wați. Avantajele bazate pe idiosincrazie ale computerelor pentru matematică, stocare și căutare, facultăți foarte puțin utilizate de strămoșii noștri, pot fi proiectate și reevaluate în laboratoare.

Facebook, Agenția Națională de Securitate și alții construiesc centre de stocare a informațiilor de ordinul exabyților, care consumă mai mult de un megawatt și se întind pe patru hectare, în timp ce ADN-ul poate stoca același volum de informații într-un miligram. În mod clar, ADN-ul nu este o tehnologie de stocare matură, dar cu Microsoft și Technicolor pariind pe el, ar fi înțelept să-i acordăm atenție. Motivul principal pentru cele șase trilioane de jouli de energie necesare pentru obținerea unei minți omenești productive sunt cei douăzeci de ani necesari instruirii.

Chiar dacă un supercomputer își poate „instrui" o clonă a sa în câteva secunde, costul energetic de a produce o clonă matură de siliciu este comparabil. Proiectarea unor genii din specia Homo Sapiens poate avea un mic impact asupra acestui proces lent, dar grăbirea dezvoltării și implantarea unei memorii extinse (ca exabyți ADN sau prin alte mijloace) poate înlocui timpul de replicare al unui computer biologic până aproape de viteza de dublare a celulelor (între unsprezece minute și douăzeci și patru de ore). Ideea este că, deși noi nu știm ce raport de bio-/homo-/

nano-/robo- hibrizi va fi dominant în fiecare etapă a evoluției noastre accelerate, putem ținti către un comportament cu grad ridicat de umanitate, corectitudine și tratament („folosire") sigur(ă) unul față de celălalt.

Declarațiile de drepturi datează încă din 1689 în Anglia. Franklin Delano Roosevelt a proclamat „Cele Patru Libertăți" – libertatea de exprimare, libertatea conștiinței, libertatea față de teamă și libertatea față de nevoi. Declarația Universală a Drepturilor Omului, emisă de Organizația Națiunilor Unite în 1948, includea dreptul la viață; interzicerea sclaviei; apărarea drepturilor atunci când sunt încălcate; libera circulație; libertatea de asociere, gândire, conștiință și religie; drepturi culturale, sociale și economice; îndatoririle persoanei față de societate; și interzicerea utilizării drepturilor în contradicție cu scopurile și principiile Națiunilor Unite.

Natura „universală" a acestor drepturi nu este universal îmbrățișată și este supusă unor critici și neconformități extinse. Cum afectează emergența inteligențelor non-Homo această discuție? Cel puțin devine din ce în ce mai dificil să te ascunzi în spatele unor intuiții vagi în ceea ce privește deciziile etice – „o să-mi dau seama când voi vedea" (spunea Potter Stewart, judecător la Curtea Supremă de Justiție a Statelor Unite în 1964) sau „înțelepciunea repugnanței" (cunoscută și ca „factorul yuck", Leon Kass, 1997), sau apeluri vagi la „bunul-simț". Deoarece avem de-a face cu minți complet străine pentru noi, am putea spune extraterestre din punctul nostru de vedere, trebuie să fim expliciți – da, chiar algoritmici.

Mașinile autonome, dronele, tranzacțiile la bursă, căutările Agenției Naționale de Securitate etc. necesită decizii rapide, aprobate în prealabil. Putem face observații interesante în multe aspecte ale eticii pe care am încercat să le izolăm și să le explicăm secole de-a rândul. Provocările au inclus priorități conflictuale, precum și predilecții cognitive întipărite la nivel biologic, sociologic și semilogic. În mod notabil departe de consens în dogme universale despre drepturile omului se află noțiunile de

confidențialitate și demnitate, chiar dacă acestea influențează multe legi și regulamente.

Oamenii ar putea avea dreptul de a citi (și schimba) mințile computerelor pentru a vedea de ce iau decizii care intră în conflict cu instinctele noastre de Homo Sapiens. Nu este just ca mașinile să ceară același lucru de la noi? Observăm creșterea mișcărilor către transparență în conflictele financiare potențiale; *software*-urile „cu sursa deschisă", *hardware* și *wetware*; Legea privind accesul echitabil la știință și tehnologie; și Open Humans Foundation.

În cartea sa din 1976, *Computer Power and Human Reason*, Joseph Weizenbaum argumenta că mașinile nu ar trebui să înlocuiască oamenii în situațiile care necesită respect, demnitate sau îngrijire, în timp ce alții (autoarea Pamela McCorduck și informaticieni precum John McCarthy și Bill Hibbard) au răspuns că mașinile pot fi mai imparțiale, calme și consecvente, mai puțin abuzive sau răuvoitoare decât oamenii în astfel de posturi.

Egalitate

La ce se referea tânărul Thomas Jefferson, în vârstă de treizeci și trei de ani, în 1776, când scria „Noi menținem aceste adevăruri a fi evidente de la sine, acela că toți oamenii sunt creați egali, acela că sunt dotați de către Creator cu anumite drepturi inalienabile și acela că printre aceste drepturi se află și dreptul la viață, dreptul la libertate și dreptul de a-și căuta fericirea"? Spectrul oamenilor din prezent este vast. În 1776, termenul „oameni" nu-i includea pe oamenii de culoare sau pe femei. Chiar și astăzi, oamenii născuți cu probleme congenitale cognitive sau comportamentale sunt destinați unui tratament inegal (fie el chiar și plin de compasiune, în unele cazuri) – sindromul Down, boala Tay-Sachs, sindromul X fragil, paralizie cerebrală și așa mai departe.

Și, pe măsură ce schimbăm locul geografic și ne maturizăm, drepturile noastre inegale se schimbă spectaculos. Embrioni, bebeluși, copii, adolescenți, adulți, pacienți, infractori, identități

de gen și preferințe de gen, cei mai bogați și cei mai săraci – toate aceste entități au drepturi și realități socioeconomice diferite. O cale către noi tipuri de minți ce obțin și păstrează drepturi similare elitelor umane ar fi să păstreze o componentă din genul Homo, cum ar fi un scut uman sau un monarh/CEO reprezentativ, ce va semna orbește documente tehnice enorme și va lua decizii financiare, de sănătate, diplomatice, militare sau de securitate pripite. Probabil ne va fi foarte greu să oprim, să modificăm sau să ștergem (să omorâm) un computer și memoriile sale – în special dacă s-a împrietenit cu oamenii și a făcut pledoarii spectaculos de convingătoare pentru supraviețuire (așa cum ar face toți cercetătorii excelenți luptând pentru viețile lor).

Chiar și Scott Adams, creatorul lui *Dilbert*, și-a spus părerea pe această temă, părere susținută de experimentele de la Universitatea Eindhoven din 2005, care observă cât de susceptibili sunt oamenii la experimentele Milgram în care robotul este victimă, făcute la Yale, începând cu anul 1961. Date fiind multele drepturi ale corporațiilor, inclusiv dreptul de a deține proprietăți, este probabil ca și alte mașini să obțină drepturi similare și va fi greu să administrăm inechitățile drepturilor selective pe gradienți multiaxiali de intelect și sentimente de înlocuire.

Reguli radical divergente pentru oameni față de nonoameni și hibrizi

Împărțirea observată mai sus pentru variația din cadrul speciei Homo Sapiens în ceea ce privește drepturile explodează într-o revoltă de inegalități imediat ce trecem la entități care se suprapun (sau se vor suprapune în curând) peste spectrul umanității. Pe Google Stree View, fețele oamenilor și numerele de înmatriculare are mașinilor sunt încețoșate. Dispozitivele video sunt excluse din multe locuri, cum ar fi curțile de judecată și ședințele comitetelor. Camerele publice și camerele care se pot purta pe corp cu softuri de recunoaștere facială intră în zona tabuurilor. Oare ar trebui

excluși și oamenii care au sindromul hipertimestic sau memorii fotografice din aceste contexte?

Nu ar trebui, oare, ca oamenii cu prosopagnozie (incapacitatea de a identifica o față cunoscută) sau uitucii să beneficieze de softuri de recunoaștere facială și recunoaștere optică a caracterelor oriunde merg și, dacă ei au acest drept, de ce nu l-ar avea oricine? Dacă avem cu toții aceste instrumente într-o anumită măsură, nu ar trebui să beneficiem cu toții de ele?

Aceste scenarii sunt un ecou al nuvelei din 1961 a lui Kurt Vonnegut „Harrison Bergeron", în care aptitudinile excepționale sunt suprimate din respect pentru cel mai mic numitor comun al mediocrității societății. Experimente mentale cum ar fi Camera chinezească a lui John Searle și Cele trei legi ale roboticii lui Isaac Asimov fac apel la tipurile de intuiții care dau bătăi de cap creierelor omenești, așa cum au demonstrat Daniel Kahneman, Amos Tversky și alții. Experimentul Camera chinezească postulează că o minte compusă din componente mecanice și Homo Sapiens nu poate fi conștientă, indiferent cât de competentă ar fi la conversația (chinezească) umană inteligentă, dacă un om nu poate identifica sursa conștiinței și nu o poate „simți". Preferința forțată pentru prima și a doua lege a lui Asimov favorizează mințile omenești în detrimentul oricărei alte minți prezente ca o umbră ștearsă în a treia lege, cea a autoconservării.

Dacă roboții nu au exact aceeași conștiință ca oamenii, atunci acest motiv este folosit ca scuză pentru a le conferi drepturi diferite, analog argumentelor conform cărora alte triburi sau rase sunt mai puțin umane. Roboții chiar manifestă liberul-arbitru? Sunt ei deja conștienți? Roboții Qbo au trecut „testul oglinzii" de autorecunoaștere și roboții NAO un test corespunzător în care și-au recunoscut propria voce, din care au dedus starea lor internă.

Pentru liberul-arbitru, avem algoritmi care nu sunt nici complet determiniști, nici aleatorii, ci au ca scop un proces probabilistic de luare a deciziilor aproape optim. Se poate argumenta că aceasta este o consecință darwinistă practică a teoriei jocurilor. În

ce privește multe (nu toate) jocuri/probleme, dacă sunt complet previzibile sau complet aleatorii, atunci avem tendința să pierdem.

Dar care este atracția liberului-arbitru? Din punct de vedere istoric, ne-a oferit o cale de a da vina în contextul recompensei și pedepsei pe Pământ sau în viața de apoi. Obiectivele pedepsei pot include ajustarea priorității indivizilor de a ajuta la supraviețuirea speciei. În cazuri extreme, asta poate însemna privarea de libertate sau alte restricții, dacă întărirea pozitivă/negativă skinneriană nu poate proteja societatea. În mod clar, astfel de instrumente se pot aplica liberului-arbitru, în sens larg – oricărei mașini al cărei comportament dorim să-l controlăm.

Putem să argumentăm dacă robotul experimentează cu adevărat *qualia* subiectivă a liberului-arbitru sau conștiinței de sine, dar același lucru se aplică și la evaluarea unui om. Cum știm dacă un sociopat, un pacient în comă, o persoană cu sindromul Williams sau un bebeluș au același liber-arbitru sau aceeași conștiință de sine ca noi? Și de ce ar conta, practic? Dacă oamenii (de orice fel) pretind în mod convingător să experimenteze conștiința, durerea, credința, fericirea, ambiția și/sau utilitatea pentru societate, ar trebui, oare, să le negăm drepturi pentru că acele *qualia* ipotetice ale lor sunt ipotetic diferite de ale noastre?

Liniile drepte și roșii ale prohibiției, peste care, teoretic, nu vom trece niciodată, par din ce în ce mai mult temporare și neclare. Linia dintre oameni și mașini nu mai este la fel de clară, atât pentru că mașinile devin mai omenești, cât și pentru că oamenii devin mai mașinali – nu doar pentru că urmăm aproape orbește script-urile GPS, tweet-urile făcute din reflex și marketingul ingenios conceput, ci și pentru că digerăm din ce în ce mai multe informații cu creierul și mecanismele noastre de programare genetică. Inițiativa NIH BRAIN dezvoltă tehnologii inovatoare și le folosește pentru a cartografia conexiunile și activitatea circuitelor mentale, astfel încât să îmbunătățească bunurile neurobiologice electronice și sintetice.

Diverse linii roșii depind de excepționalismul genetic, în care genetica este considerată o moștenire permanentă (deși se poate

dovedi că este reversibilă), în timp ce tehnologiile scutite (și letale) de la aceasta, cum ar fi automobilele, sunt practic ireversibile din cauza forțelor sociale și economice. În cadrul geneticii, o linie roșie ne face să interzicem sau să evităm alimentele modificate genetic, dar să îmbrățișăm bacteriile modificate genetic care creează insulină sau oamenii modificați genetic – martori stau tratamentele aprobate în Europa pentru adulți și embrioni umani.

Linia pentru manipularea germenilor pare mai puțin sensibilă decât media, o linie practică trasă între siguranță și eficiență. Căsătoriile între doi transportatori sănătoși ai aceleiași boli genetice au de ales între a nu face copii, 25% șanse de a-și pierde embrionul prin avort (spontan sau indus), 80% șanse de pierdere prin fertilizare in vitro sau, posibil, o pierdere potențială de 0% a embrionului prin inginerie genetică a spermei (linia germinală). Pare prematur să declarăm această ultimă opțiune improbabilă.

Pentru cercetările „pe oameni", consultăm Declarația de la Helsinki, din 1964, aducându-ne aminte și de experimentul cu sifilis Tuskegee 1932–72, posibil cea mai notorie cercetare biomedicală din istoria Statelor Unite. În 2015, Nonhuman Rights Project a intentat un proces la Curtea Supremă din New York în numele a doi cimpanzei păstrați pentru cercetare de Universitatea Stony Brook. Decizia curții de apel a fost că cimpanzeii nu trebuie tratați ca niște persoane din punct de vedere legal pentru că „nu au îndatoriri și responsabilități în societate", în ciuda a ceea ce susțineau Jane Goodall și alții că, de fapt, ar avea, precum și în ciuda argumentelor că o astfel de decizie ar putea fi aplicată și copiilor sau persoanelor cu handicap.*

Ce împiedică extrapolarea acestei decizii și asupra altor animale, organoide, mașini și hibrizi? Promovând interdicții asupra „armelor autonome" (Hawking, Musk, Tallinn, Wilczek, Tegmark și alții), am demonizat un tip de mașină „neinteligentă", în timp

* https://www.nbcnews.com/news/us-news/lawyer-denying-chimpanzees-rights-could-backfire-disabled-n734566 (n.a.).

ce alte mașini – de exemplu, cele compuse din mulți Homo Sapiens care votează – pot fi mai letale și mai prost îndrumate.

Transoamenii au populat deja Pământul? Să ne gândim la populațiile care nu au avut contact cu alți oameni, cum ar fi Sentinelese și Andamanese din India, Korowai din Indonezia, Mashco-Piro din Peru, Pintupi din Australia, Surma din Etiopia, Ruc din Vietnam, Ayoreo-Totobiegosode din Paraguay, Himba din Namibia și zeci de triburi din Papua-Noua Guinee. Cum ar răspunde acestea sau strămoșii noștri? Am putea defini ca „transumani" oamenii și culturile pe care oamenii care trăiesc într-o cultură modernă, însă care nu este tehnologizată, nu i-ar putea înțelege.

Astfel de oameni moderni din Epoca de Piatră ar înțelege foarte greu de ce sărbătorim dovezile recente ale undei gravitaționale LIGO care susțin teoria generală a relativității veche de o sută de ani. S-ar scărpina în cap gândindu-se de ce avem ceasuri atomice sau sateliți GPS pentru a putea ajunge acasă sau cum și de ce ne-am extins vederea de la o bandă optică restrânsă la spectrul complet de la radio la gamma. Ne putem mișca mai repede decât orice altă specie în viață; putem atinge viteza necesară pentru a scăpa de gravitația Pământului și putem supraviețui în vidul rece al spațiului.

Dacă aceste caracteristici (și alte câteva sute) nu constituie transumanism, atunci ce? Dacă simțim că judecătorul transumanismului nu ar trebui să fie oameni paleoculturali, ci oamenii recenți, atunci cum putem ajunge la transumanism? Noi, „oamenii recenți", vom putea întotdeauna să înțelegem fiecare increment tehnologic nou – nefiind niciodată surprinși îndeajuns pentru a declara atingerea unei ținte transumane (în mișcare). Profetul science-fiction William Gibson spunea: „Viitorul este deja aici – doar că nu este distribuit uniform". Deși acest lucru subestimează următoarea rundă a „viitorului", cu siguranță milioane de oameni sunt deja transumani – majoritatea dintre noi dorind și mai mult. Întrebarea „Ce a fost un om?" s-a metamorfozat deja în „Care au fost multele tipuri de transoameni?... Și care au fost drepturile lor?"

Capitolul 24
Utilizarea artistică a ființelor cibernetice

Caroline A. Jones

Caroline A. Jones este profesor de istoria artei la Departamentul de Arhitectură de la MIT și autoarea cărților *Eyesight Alone: Clement Greenberg's Modernism Bureaucratization of the Senses, Machine in the Studio: Constructing the Postwar American Artist* și *The Global Work of Art.*

Interesul lui Caroline A. Jones în arta modernă și contemporană este îmbogățit de disponibilitatea de a se cufunda în tehnologiile implicate în producerea, distribuirea și recepția ei. „În calitate de istoric al artei, multe dintre întrebările mele se referă la ce fel de artă putem face, ce tip de gânduri putem produce și cel fel de idei putem găsi care-l pot împinge pe om dincolo de parametrii noștri încăpățânați, egoiști, «concentrați doar pe grupul nostru mic». Filosofii și filosofiile de care sunt atrasă sunt cei care pun sub semnul întrebării obsesia occidentală a individualismului. Aceștia vin din foarte multe locuri diferite și reînvie atât de multe întrebări și probleme care au fost puse în anii 1960."

Recent, și-a concentrat atenția pe istoria ciberneticii. Cursul său de la MIT – Mașini automate, automatism, sisteme, cibernetică – explorează istoria interfeței dintre om și mașină în termeni de *feedback*, explorând latura culturală, și nu pe cea inginerească, a acestei idei. Începe cu lecturi obligatorii din Wiener, Shannon și Turing și apoi pivotează de la oamenii de știință și ingineri la lucrările și ideile artiștilor, feministelor și teoreticienilor postmoderni. Scopul ei: să ajungă la o paradigmă centrală nouă a evoluției care este bazată pe cultură – „comunalismul și simbioza dintre specii, mai degrabă decât supraviețuirea celui mai bine adaptat individ".

În calitate de istoric, Caroline face diferența între ceea ce ea a denumit „cibernetica de stânga" și „cibernetica de dreapta": „Ce înseamnă cibernetica de stânga? La un anumit nivel, este un joc de cuvinte: cibernetica «lăsată» în urmă*. La alt nivel, este o grupare politică vagă ce denotă coasta stângă: California, Esalen, grupul pe care Dave Kaiser îl numește «fizicienii hipioți». Nu este un termen corespunzător, dar este o cale de a recunoaște că a existat un grup controlat de complexul industrial militar, uneori în mod nefericit, care ne-a oferit instrumentele critice".

* În limba engleză, „left" are sensul de „stânga", dar și de „lăsat" sau „rămas" (n.t.).

> Arta cibernetizată este foarte importantă, dar arta pentru viață cibernetizată este mai importantă.
>
> — *NAM JUNE PAIK, 1966*

Inteligența artificială nu este ce și-au dorit inițial artiștii de la cibernetică, atunci când a apărut cartea lui Norbert Wiener, *The Human Use of Human Beings: Cybernetics and Society*, în anul 1950. Gama de artiști care se identificau cu cibernetica în anii 1950 și 1960 a avut inițial puțin acces la „mașini de gândit". În plus, ingineri iscusiți făcuseră deja roboți țestoase, jongleri și roboței care trăgeau la lumină, dar nu și creiere gigantice. Folosind conexiuni, fire de cupru, comutatoare simple și senzori electronici, artiștii au mers pe urmele ciberneticienilor, creând sculpturi și medii care simulau o conștiință interactivă – mișcări analoge și interfețe care aveau mai mult de-a face cu instinctele și politicile sexuale de după război decât cu automatizarea producției de cunoștințe. Obscurizată acum de o ideologie a „inteligenței" libere, nepriponite de componente *hardware* sau de trup, inteligența artificială a uitat de primele zile ale asimilării ciberneticii de către artiști. Acele eforturi merită revizitate; au modelat relațiile cu ceea ce filosofii francezi Gilles Deleuze și Félix Guattari au denumit „*phylum*-ul mașinilor", care are de-a face cu modul în care oamenii gândesc și simt în corpurile conectate la o lume fizică, materială, stimulantă emoțional și semnalizatoare.

Cibernetica se pare că s-a prăbușit acum într-un discurs atotpătrunzător despre inteligența artificială care era departe de a fi predestinată. Cuvântul „cibernetică" a căpătat o prospețime nouă pentru concepte care aveau cel puțin patru secole: noțiuni despre *feedback*, amortizare automată, homeostază biologică, calcul logic și gândire sistemică ce au existat încă din Iluminism (fiind sporite de Revoluția industrială). Numele din această descendență îi includ pe Descartes, Leibniz, Sadi Carnot, Clausius, Maxwell și Watt. Însă numele dat de Wiener a avut efecte culturale profunde. Ubicuitatea în zilele noastre a prefixului *ciber-* confirmă dorința de semnificant nou pentru relațiile întortocheate dintre oameni și mașini. Cu sensul pe care l-a folosit Wiener, lucrurile „ciber" implicau pur și simplu „control și comunicare între animal și mașină". Dar, după revoluția digitală, „ciber" a trecut dincolo de servomecanisme, cicluri de *feedback* și comutatoare, înglobând noțiuni de *software*, algoritmi și cyborgi. Activitatea artiștilor înclinați către cibernetică privește comportamentele emergente ale vieții care scapă inteligenței artificiale în condiția sa curentă.

În ceea ce privește sensul inițial, Wiener a ajuns până la greaca antică pentru a împrumuta cuvântul pentru „timonier" (κυβερνήτης/*kubernétés*), o figură masculină care canalizează putere și instinct la timona navei, care citea valurile, judeca vântul, ținea o mână pe cârmă și îi direcționa pe sclavi când trăgeau fără discernământ (mecanic) la vâsle. Cuvintele grecești migraseră deja în engleza modernă prin latină, trecând de la *kuber-* la *guber-*, rădăcina cuvintelor „guvernamental" și „guvernator", alt termen pentru controlul masculin, dezvoltat de James Watt pentru a-și descrie dispozitivul său din secolul al XIX-lea pentru modularea unui motor cu aburi impetuos. Cibernetica a preluat astfel idei care făceau de mult timp analogii între oameni și dispozitive și le-a generalizat într-o știință aplicată, adăugând particula „-ică". Cei trei „*c*" ai lui Wiener (comandă, control, comunicare) se bazau pe matematica probabilității pentru a formaliza sisteme (biologice sau mecanice) teoretizate ca un set de intrări de informații care obține ieșiri sub formă de acțiune într-un mediu – o agendă

musculoasă și trupeșă, adesea minimalizată în genealogiile inteligenței artificiale.

Dar etimologia nu reușește să exprime suficient entuziasmul resimțit de participanți, când matematica s-a unit cu biologia teoretică (Arturo Rosenblueth) și teoria informației (Claude Shannon, Walter Pitts, Warren McCulloch), pentru a produce un baraj de cercetări și publicații interdisciplinare văzute ca prezentând posibilitatea de a schimba nu doar modul în care se face știință, ci și modul în care oamenii din viitor vor interacționa cu tehnosfera. În cuvintele lui Wiener, „ne-am modificat mediul atât de radical încât trebuie acum să ne modificăm pe noi înșine pentru a putea exista"*. Întrebarea presantă este: cum ne modificăm? Mergem în direcția potrivită sau ne-am pierdut calea, devenind instrumentele instrumentelor noastre? Revizitarea istoriei recente a contribuției umaniștilor/artiștilor la cibernetică ne-ar putea ajuta să ne orientăm către un viitor mai puțin periculos și mai etic.

Anul 1968 a fost un vârf al difuziei culturale și al asimilării artistice a termenului. În acel an, galeria Howard Wise și-a deschis expoziția de „sculptură cibernetică" a lui Wen-Ying Tsai în centrul Manhattanului, iar emigranta politică poloneză Jasia Reichardt și-a deschis expoziția „Serendipitatea cibernetică" la Institutul de Artă Contemporană din Londra (cuvântul „cibernetic" din titlul ei voia să însemne „făcut cu sau de un computer", chiar dacă majoritatea operelor de artă expuse nu aveau computere concrete în circuitele lor). Cele două decenii dintre 1948 și 1968 au reprezentat atât răspândirea conceptelor cibernetice într-o cultură mai largă, cât și răspândirea mașinilor de calcul într-o migrare lentă, de la echipamente militare brevetate până la corporațiile multinaționale și laboratoarele universitare, unde artiștii au început să primească acces. Disponibilitatea componentelor cibernetice – „organe senzoriale" (ochi electronici,

* *The Human Use of Human Beings* (Boston: Houghton Mifflin, 1954), 46 (n.a.).

senzori de mișcare, microfoane) și „organe făptuitoare" (conexiuni electronice, comutatoare, sisteme hidraulice și pneumatice) – pe frontul hobby-urilor de acasă făcea din computer mai puțin un „creier electronic" decât un organ ajutător dintr-un set de componente. Nu exista încă o metaforă predominantă pentru „inteligența artificială". Așadar artiștii erau deja bricoleuri de corpuri electronice, interesați mai mult de acțiuni decât de calcule sau cogniție. Existau indicii care făceau din „computere" calculatoare în cursa către *Homo rationalis*, dar mai mult în aspirație decât în realizare.

În lumina convergenței digitale de astăzi a instrumentelor de imagistică artistică/științifică, spectacolul lui Reichardt era profetic în insistența sa în ceea ce privește confundarea granițelor dintre artă și ceea ce ar putea fi numită „știință creativă aplicată". Conform catalogului, „niciun vizitator al expoziției, dacă nu citește toate notele referitoare la toate lucrările, nu va ști dacă privește ceva făcut de un artist, inginer, matematician sau arhitect". Astfel, robotul hilar de disfuncțional al lui Nam June Paik, *Robot K–456* (1964), prezentat pe coperta catalogului și descris ca „un robot feminin cunoscut pentru comportamentul său deranjant și idiosincratic", se va întrece cu un *Colloquy of Mobiles* (1968), coregrafic construit de ciberneticianul mai puțin cunoscut Gordon Pask. Pask a lucrat cu un scenograf de teatru londonez pentru a construi un aparat „masculin" fusiform cu balamale și tije, configurat să comunice cu entitățile „feminine" bulboase din fibră de sticlă care se aflau prin preajmă. Dacă cineva ar putea să cartografieze caracteristicile programului (sau să spicuiască teatrul său reacționar de gen) fără a citi eseul din catalog este o întrebare deschisă. Semnificativă este concentrarea lui Pask pe comportamentele automatelor sale, pe interactivitatea lor, pe responsivitatea lor într-un mediu modulat artificial și pe „reflectarea" comportamentului uman.

„Serendipitatea cibernetică" de la Institutul de Arte Contemporane a introdus o paradigmă importantă: ecosistemul mașinic, în care privitorul este o parte biologică, având ca sarcină să-și dea

seama care sunt declanșatorii interacțiunii. Vizitatorii din acele galerii londoneze au devenit brusc „organisme cibernetice" – cyborgi –, deoarece, pentru a experimenta corespunzător arta, trebuia să intri într-o conversație simbiotică cu servomecanismele. Această orientare către mediile interactive dintre om și mașină ca estetică devine din ce în ce mai clară când examinăm câteva alte opere de artă ale perioadei, începând cu una care constituie un exemplu timpuriu al comportamentului emergent – *Senster*, sculptura interactivă făcută de artistul/inginerul Edward Ihnatowicz (1970), celebrată de inginerul în robotică medicală Alex Zivanovic, editor al site-ului web dedicat carierei puțin cunoscute a lui Ihnatowicz, ca „una dintre primele opere de artă robotice interactive controlate de computer". Aici „computerul" își face intrările sale (deși este un dispozitiv limitat, cu doisprezece biți). Dar mai degrabă decât „inteligența", Ihnatowicz a urmărit un avatar al comportamentului afectiv. Cheia succesului neașteptat al lui *Senster* a fost programarea cu care Ihnatowicz făcea aparatul hidraulic de cinci metri (designul său cu balamale și aspect de miraj era inspirat de un clește de homar) să transmită sfială când răspundea oamenilor din proximitate. Canalele de sunet și senzorii de mișcare ai lui *Senster* erau setați să se tragă în spate la zgomote puternice sau mișcări bruște, aparent agresive. Doar oamenii care erau dispuși să vorbească blând și să-și moduleze gesturile urmau să fie recompensați de abordarea liniștită și curioasă a lui *Senster* – o experiență care a devenit reală și pentru Ihnatowicz însuși, când acesta a structurat inițial programul, iar mașina s-a întors către el cu solicitudine, după ce a tușit pentru a-și elibera gâtul.

În aceste utilizări artistice ale ființelor cibernetice, simțim o necesitate crescândă de a instrui publicul în a se resimți pe sine ca integrat într-un mediu tehnologizat, modificându-se pe sine pentru a comunica intuitiv cu mașinile. Această necesitate a devenit deja explicită în expoziția de „Sculptură Cibernetică" a lui Tsai. Cei care experimentează instalația lui în care se cufundă cu totul se așteptau să experimenteze cu viața mașinică: ce

comportament ar declanșa servomecanismele? Asemenea, participantul din galeria umană ar fi trebuit să explice protocolul: „Bate din palme – asta face sculpturile să răspundă". În cuvintele unui critic de atunci:

> Un schelet de tije din oțel se ridică de pe o placă. Această bază vibrează cu treizeci de cicluri pe secundă; tijele flexează rapid, în curbe armonice. Amplasat într-o cameră întunecată, este iluminat de lumini stroboscopice. Pulsul luminilor intermitente variază – sunt conectate la senzori de sunet și proximitate. Rezultatul este că, atunci când cineva se apropie de un Tsai sau face un zgomot în vecinătatea lui, obiectul răspunde. Tijele par să se miște; se observă o sclipire, o lumină și un balet straniu de metal, ale cărui mișcări variază de la nemișcat la tremurat și apoi înapoi la o ondulație lentă, inefabil de senzuală.

La fel ca *Senster*, aparatul stimula (și simula) o interacțiune afectivă, și nu una rațională. Oamenii simțeau că întâlnesc comportamente ce indică viața sensibilă; entitățile lui Tsai era adesea clasificate ca „vegetale" sau „acvatice". Astfel de ambiții ce țin de mediu și de mișcare erau notorii în lumea internațională a artei din vremea aceea. Dincolo de grajdul de la Howard Wise, erau emigranții politic care formau colectivul GRAV în Paris, „arhitecturile cibernetice" ale lui Nicolas Schöffer, giratiile de lumină și plastic ale German Zero Gruppe și altele – toate definind și informând genul de artă cibernetică ce se anunța.

Utilizarea artistică a ființelor cibernetice de la sfârșitul anilor șaizeci nu a făcut nicio investiție în „inteligență". Știind că mașinile sunt lipsite de inteligență și incapabile de emoție, acești creatori erau încrezători când puneau în scenă simulări oneste. Ce-i interesa pe ei erau acțiuni, instincte și afecte ce evocau mișcări ale mașinii; imitau comportamente sexuale și animalice, ca și cum ar fi sub pragul conștiinței. Astfel de artiști nu erau interesați de manipularea datelor sau informațiilor (deși Hans

Haacke se va fi îndreptat în acea direcție în 1972 cu lucrările sale de „sisteme în timp real"). Cultura cibernetică pe care artiștii și cercetătorii o configurau pe două continente integra omul în tehnosferă și seducea percepția cu comportamentele grațioase și sensibile ale filonului mașinic. „Artificialul" și „naturalul" erau întrețesute în această estetică cibernetică timpurie.

Dar nu se va termina aici. Crucială pentru extinderea acestui set necritic și în mare parte masculin de medii cibernetice se va dovedi a fi o cohortă critică radicală de artiste uimitoare care vor apărea în anii 1990, pe deplin conștiente de predecesorii lor în artă și tehnologie, dar poate mai mult inspirate de fondatoarele feministe ale revistei din 1970 *Radical Software* și boomul cultural al inspiratei polemici din 1984 a Donnei Haraway, „Manifestul cyborgului". Teatrul scârțâitor de gen al lui Paik și Pask, creaturile inocente ale lui Ihnatowicz și Tsai au fost mobilizate ca istețe, performative și postmoderne, ca în lecția lui Lynn Hershman „Dollie Clone Series" (1995–1998), care consta în asamblajele interactive *CybeRoberta* și *Tillie, păpușa telerobotică*, lucrând în tehnosferă cu profesionalismul burlescului, făcându-ne cu ochiul și pliindu-ne pe noi, privitorii, într-o conștiință explicită a poziției noastre voyeuriste, atât în calitate de subiecți care privesc, cât și ca obiecte de privit.

Tehnosfera „inocentă" stabilită de sculptorii cibernetici bărbați din anii 1960 a fost, până în anii 1990, identificată de artistele feministe drept o condiție complet înecătoare ce ne solicită atenția critică. În același timp, feministele au abordat problema cărei „inteligențe" încearcă să o simuleze inteligența artificială. Pentru o artistă ca Hershman Leeson, răspunzând la „triumful" tehnic de a o clona pe oaia Dolly, era crucial să facem conexiunea între producția de carne și „mașinile de carne". Hershman Leeson producea „păpuși" drept clone, oferind o încadrare critică a modului în care individualizarea contemporană a devenit o parte din domeniul logic, replicativ și plastic.

Deși tehnofeministele din anii 1990 și 2000 nu erau „ciber" tot timpul, lucrările lor au complicat calitățile dominante mașinice

și cinetice ale mediilor tehnologice anterioare ale artiștilor bărbați. Telecyborgul androgin din *Imagination, Dead Imagine* (1991) al lui Judith Barry, de exemplu, nu avea nicio parte în mișcare. El/ea era compus din semnale pure, proiecții intermitente pe suprafețe plate. În lucrarea sa, Barry a comentat asupra efectelor alienante ale tehnologiei secolului XX. Imaginea unui cap androgin umple un cub enorm făcut din ecrane de trei metri pe cinci părți, montate pe o bază de oglinzi tot de trei metri. O diversitate de lichide vâscoase și care nu arată prea bine (galbene, roșietice-portocalii, maro), materiale uscate (rumeguș? făină?) și chiar insecte își croiesc drum pe cap, a cărui sublimitate stoică este minunat virtualizată pe ecranele enorme ale lucrării. *Dead Imagine*, prin forma sa „platonică" la scară mare și cubică, rămâne atât artificială, cât și blocată în corp – refuzând o „inteligență" detașată ca nefiind deloc inteligență.

Artiștii din noul mileniu moștenesc această tradiție critică și locuiesc în paradigmele curente ale inteligenței artificiale, care au alunecat de la simulări parțiale la pretenții de inteligență. În propunerea din 1955, care este considerată a fi prima utilizare tipărită a sintagmei „inteligență artificială", informaticianul John McCarthy și colegii săi Marvin Minsky, Nathaniel Rochester și Claude Shannon au emis ipoteza că „toate aspectele învățării sau orice altă caracteristică a inteligenței pot fi, în principiu, atât de precis descrise încât o mașină le poate simula". Acest obiectiv teoretic modest s-a umflat în ultimii șaizeci și patru de ani, fiind exprimat acum de Google DeepMind ca o ambiție de a „rezolva inteligența". De a sparge codul! Dar, din păcate, ce auzim crăpând nu este codul, ci capitalismul la scară mică, contractul social și structura politeții. Depășirea meseriilor de șofer de taxi și de camion, robotizarea marketingului direct, hegemonizarea divertismentului, privatizarea utilităților și depersonalizarea îngrijirii medicale – sunt acestea „bicele" de care Wiener se temea că vom învăța să le iubim?

Artiștii nu pot rezolva niciuna dintre aceste probleme. Dar ne pot reaminti de potențialul creativ al căilor nebătătorite – intersecțiile

care ne ieșeau în cale în anii 1970, înainte ca „informația" să devină capital și „inteligența" să fie echivalată cu colectarea de date. Puternic evocator a ce se poate face cu instrumentele contemporane când revizităm posibilitățile din trecut este „licuriciul" artistului francez Philippe Parreno, numit astfel pentru a evita repetiția numelui său adevărat: *Cu un comportament instinctiv ritmic pentru a putea să călătorești dincolo de forțele existente ale vieții* (2014). Descrisă de artist ca un „automat", instalația sculpturală juxtapune o proiecție intermitentă cu desene alb-negru de licurici cu o bandă de figuri binare oscilante, având culoarea verde pe negru. Desenele pe figuri binare sunt animate folosind algoritmii matematicianului John Horton Conway, folosiți în Jocul Vieții din 1970, „automatul celular".

Conway a configurat parametrii astfel încât orice pătrat („celulă") să fie aprins („viu") sau stins („mort") într-o matrice bidimensională infinită. Regulile sunt rezumate după cum urmează: o celulă individuală va muri rapid de singurătate. Dar o celulă care atinge trei sau mai multe celule „vii" va muri și ea „din cauza aglomerației". O celulă supraviețuiește și prosperă dacă are doar doi vecini... și așa mai departe. Când o celulă moare, poate crea condițiile necesare supraviețuirii altor celule, creând tipare care par să se miște și să crească, aprinzându-se în matrice ca impulsurile neuronale evanescente sau grupuri bioluminiscente de diatomee. În filmul din 2012 al lui Stephen Hawking *The Meaning of Life*, naratorul descrie modelul matematic al lui Conway ca simulând „cum ceva complex așa cum este mintea ar putea lua naștere dintr-un set elementar de reguli", revelând ambițiile înfumurate ce caracterizează inteligența artificială contemporană: „Aceste proprietăți complexe iau naștere din legi simple ce nu conțin concepte cum ar fi mișcarea sau reproducerea", însă produc „specii", și celulele „se pot chiar reproduce, la fel cum face viața în lumea reală".*

* Narration in Stephen Hawking's *The Meaning of Life* (Smithson Productions, Discovery Channel, 2012) (n.a.).

La fel cum face viața? Artiștii cunosc lingușirile simulărilor și reprezentării, diferența dintre geniul artificiului și realitățile a „ceea ce face viața". Lucrarea lui Parreno este un ansamblu intuitiv pentru modul în care experimentăm noi „viața" prin angajamente încorporate, perspectiviste. Conștiința noastră este întrețesută electric (cibernetic), însă nu răspundem ca și cum acest set de simulări elegante generat de oameni ar avea propria sa *inteligență*.

Utilizarea artistică a ființelor cibernetice ne reamintește, de asemenea, că și conștiința însăși nu este doar „aici". Intră și iese, armonizând aceste semnale senzoriale sclipitoare. Mintea se întâmplă mult în afara limitelor craniului (și simulacrului acestuia, „placa de bază"). Cum spunea Mary Catherine Bateson, parafrazând conceptul tatălui său, Gregory, de cibernetică de rangul doi, mintea este un material „nu neapărat definit de o limită precum o bășică de piele"*. Parreno împerechează simulările artei cu simulările matematicii pentru a forța ideea de sorginte wieneriană că un astfel de model nu este, în sine, la fel ca viața. Modele sunt doar părți ale sistemelor de semnalizare ce constituie „inteligență" doar atunci când omoloagele lor biologice interacționează cu ele în creare de sens. Inteligența artificială contemporană s-a marginalizat instrumentalizând și particularizând sarcini și subrutine, confundând aceste exerciții cu înțelepciunea veritabilă. Scurta istorie culturală oferită aici ne reamintește că a privi datele ca pe o formă de inteligență, rețelele digitale ca „neuronale" sau indivizii izolați ca unități de viață este o metodă străină chiar și pentru simulările brute ale lui Conway.

Putem stigmatiza aroganța căpoasă a inteligenței artificiale curente precum „cibernetica de dreapta", calea care a condus la sistemele de armament automatizate pe care le avem în prezent, ostilitatea prost deghizată a companiei Uber pentru șoferii umani și visurile capitaliste ale companiei Google. Acum trebuie să ne întoarcem la cibernetica de stânga – teoreticieni biologi și

* Mary Catherine Bateson, 1999 foreword to Gregory Bateson, *Steps to an Ecology of Mind* (Chicago: University of Chicago Press, 1972), xi (n.a.).

antropologi încercând să înțeleagă sistemele inteligente independent de specie. Observația lui Gregory Bateson conform căreia corporațiile nu fac decât să simuleze „agregate de părți de persoane", cu decizii de maximizare a profitului tăiate din „părți ale minții mai mari și mai înțelepte", nu a fost niciodată mai oportună.*

Epistemologia cibernetică oferită aici sugerează o abordare nouă. Mintea individuală este imanentă, nu doar pentru corp, dar și pentru căi din afara corpului, și există o Minte mai mare, pentru care mintea individuală este doar un subsistem. Această Minte mai mare, susține Bateson, este comparabilă cu Dumnezeu și este, probabil, ceea ce unii oameni înțeleg prin conceptul de Dumnezeu, dar tot este imanentă sistemului social total interconectat și ecologiei planetare. Aceasta nu este iluzia colectivă a unui „Dumnezeu" exterior, care vorbește din afara conștiinței umane (această idee cu tradiție lungă de monoteism, sugerează Bateson, conduce la viziuni despre natură și mediu ca fiind, de asemenea, în afara omului „individual", făcând din ele „daruri bune de exploatat"). Mai degrabă, „Dumnezeul" lui Bateson este un locțiitor al experienței noastre evanescente de interacțiune a conștiinței în lume: Mintea mai mare, ca rezultat al intrărilor și acțiunilor care devin apoi intrări pentru alte acțiuni împreună cu alte entități – pânze de relații simbiotice ce formează tipare cu care trebuie să armonizăm și să simțim urgent.**

De la Tsai în 1970 până la Hershman Leeson în 1990 și Parreno în 2014, artiștii au criticat cibernetica de dreapta și s-au jucat cu experimente cibernetice alternative, întrupate, de mediu ale inteligenței „artificiale". Utilizarea lor artistică a ființelor cibernetice oferă înțelepciunea simbioților experimentată în tipurile de *poiesis* care pot fi obținute în această lume: ritmuri de semnale și acțiuni intuitive ce produc mișcările vieții în parteneriat cu tehnosfera electromecanică și electromagnetică. Viața, în misterioasele sale îmbinări negentropice cu materia și Mintea.

---

* *Steps to an Ecology of Mind*, 452 (n.a.).

** *Steps to an Ecology of Mind*, 467–68 (n.a.).

Capitolul 25
Inteligența artificială și viitorul civilizației

Stephen Wolfram

Stephen Wolfram este om de știință, inventator, fondatorul și CEO-ul companiei Wolfram Research. Este creatorul programului de tehnică de calcul simbolică Mathematica și al limbajului său de programare, limbajul Wolfram, precum și al motorului de cunoștințe Wolfram|Alpha. De asemenea, este autorul cărții *A New Kind of Science*. În cele ce urmează, veți găsi o transcriere editată dintr-un interviu live cu el realizat în decembrie 2015.

În aproape patru decenii, Stephen Wolfram a fost un pionier în dezvoltarea și aplicarea gândirii computaționale, fiind responsabil de multe inovații în știință, tehnologie și afaceri.

Lucrarea sa din anul 1982 *Mașinile automate celulare ca sisteme simple cu autoorganizare*, scrisă la vârsta de douăzeci și trei de ani, a fost prima dintre numeroasele sale contribuții științifice semnificative care au avut ca scop înțelegerea originilor complexității în natură.

În această perioadă, Stephen a intrat o scurtă perioadă în viața mea. Tocmai înființasem The Reality Club, o adunare neoficială de intelectuali, care s-au întâlnit în New

York pentru a-și prezenta lucrările în fața colegilor de breaslă din alte discipline. (Observație: în 1996, The Reality Club a ajuns online pe site-ul Edge.org.) Primul nostru vorbitor? Stephen Wolfram, un „wunderkind“* care a ajuns la Princeton, la Institutul de Studii Avansate. Îmi amintesc maniera sa concentrată când s-a așezat pe o canapea din sufrageria mea și a vorbit fără întrerupere aproximativ o oră în fața grupului strâns.

Din acel moment, Stephen are ca scop să facă toate cunoștințele lumii ușor computabile și accesibile. Programul său, Mathematica, este sistemul definitoriu pentru tehnica de calcul modernă. Wolfram|Alpha calculează răspunsuri la nivel de expert utilizând tehnologia inteligenței artificiale. Consideră că limbajul său de programare Wolfram este primul limbaj adevărat de comunicare computațională pentru oameni și inteligențe artificiale.

M-am reîntâlnit cu el cu patru ani în urmă, în Cambridge, Massachusetts, pentru o discuție liberă pe tema inteligenței artificiale. Stephen a intrat, m-a salutat, s-a așezat și, privind camera video aranjată ca să înregistreze conversația pentru Edge, a început să vorbească și nu s-a oprit timp de două ore și jumătate.

Eseul care urmează este o versiune editată a acelei sesiuni, care a fost un fel de curs aprofundat despre Wolfram, fiind un mod corespunzător de a încheia acest volum – la fel cum prelegerea lui Stephen de la Reality Club din anii 1980 a fost un mod minunat de a iniția întreprinderea intelectuală în desfășurare al cărei rezultat este comunitatea bogată de gânditori care-și prezintă activitatea unul altuia și publicului în această carte.

---

* „Copil-minune“ în limba germană (n.t.).

Eu văd tehnologia ca fiind automatizarea obiectivelor omenești pentru a le face executabile pentru mașini. Obiectivele omenești din trecut au presupus mutarea obiectelor dintr-un loc în altul cu ajutorul unui stivuitor, și nu al mâinilor. Acum munca pe care o putem face automat, cu mașinile, este mai mult mentală decât fizică. Este evident că putem automatiza multe dintre sarcinile pe care noi, oamenii, ne mândrim de mult timp că le-am făcut singuri. Care este viitorul condiției umane în această situație?

Oamenii vorbesc despre viitorul mașinilor inteligente și dacă vor prelua controlul și vor hotărî ce să facă ele însele. Dar inventarea obiectivelor nu este ceva ce are o cale în automatizare. Cineva sau ceva trebuie să definească aceste obiective ale unei mașini – ce încearcă să execute. Cum sunt definite obiectivele? Pentru om, au tendința să fie definite de istoria personală, de mediul cultural, de istoria civilizației noastre. Obiectivele sunt unic omenești. În ceea ce privește mașina, noi îi conferim un scop atunci când o construim.

Ce fel de lucruri posedă inteligență, obiective sau scopuri? Momentan, cunoaștem un exemplu mare și lat, și anume noi – creierele noastre, inteligența noastră omenească. Inteligența omenească, presupuneam eu cândva demult, este mult peste orice altceva există natural în lume; este rezultatul unui proces elaborat de evoluție, evidențiindu-se astfel de restul existenței. Dar am realizat, ca urmare a științei pe care am practicat-o, că nu așa stau lucrurile.

Oamenii pot spune, de exemplu, „Vremea are propria ei minte". Aceasta este o afirmație animistă și pare să nu aibă ce căuta în gândirea științifică modernă. Dar nu este atât de neserioasă pe cât sună. Ce face creierul omenesc? Un creier primește anumite intrări senzoriale, calculează, determină înfăptuirea anumitor acțiuni și generează un anumit rezultat. La fel ca vremea. Tot soiul de sisteme, de fapt, calculează – indiferent dacă este vorba despre un creier sau, să spunem, un nor care răspunde la mediul său termic.

Putem să argumentăm că facem calcule mult mai sofisticate cu creierele noastre decât cele care apar în atmosferă. Dar se pare că există o echivalență mare între tipurile de calcule pe care le fac diversele tipuri de sisteme. Acest lucru pune într-o nouă lumină problema condiției umane, pentru că se pare că nu suntem atât de speciali pe cât credeam. Sunt toate aceste sisteme ale naturii care sunt, practic, echivalente în ceea ce privește capacitățile de calcul.

Ce ne face diferiți de toate celelalte sisteme sunt elementele particulare ale istoriei noastre, care ne dau noțiunile pe care le avem noi de scop și obiectiv. Acesta este un alt fel de a spune că, deși cutia de pe biroul nostru gândește la fel de bine precum creierele noastre, nu va avea niciodată, intrinsec, obiective și scopuri. Acestea sunt definite de trăsăturile noastre particulare – de biologia, psihologia și istoria noastră culturală.

Când ne gândim la viitorul inteligenței artificiale, trebuie să ne gândim la obiective. Cu ele contribuie oamenii; cu ele contribuie civilizația noastră. Execuția acestor obiective este ceea ce putem automatiza din ce în ce mai bine. Care va fi viitorul oamenilor într-o astfel de lume? Ce rămâne de făcut atunci? Unul dintre proiectele mele a fost să înțeleg evoluția scopurilor omenești în timp. Astăzi avem tot felul de scopuri. Dacă privim înapoi cu o mie de ani, oamenii aveau cu totul alte obiective: cum fac rost de mâncare? Cum mă apăr? În lumea occidentală modernă, în cea mai mare parte, nu petrecem prea mult timp din viața noastră gândindu-ne la aceste scopuri. Din punctul de vedere al unui om de acum o mie de ani, unele obiective pe care le au

oamenii din ziua de astăzi ar părea complet bizare – de exemplu, alergatul pe bandă. Acum o mie de ani, ar fi fost nebunie curată.

Ce vor face oamenii în viitor? Multe dintre scopurile pe care le avem astăzi sunt generate de un anumit tip de raritate. Resursele din lume sunt limitate. Oamenii vor să obțină mai multe dintr-o anumită categorie. Până și timpul este puțin în viața noastră. În cele din urmă, acele forme de raritate vor dispărea. Cea mai pregnantă discontinuitate va fi cu siguranță când vom obține efectiv imortalitatea omenească. Dacă aceasta va fi obținută biologic sau digital nu știm încă, dar o vom obține inevitabil. Multe dintre obiectivele noastre au la bază cel, puțin parțial, mortalitatea noastră: „Voi trăi doar o perioadă de timp, așa că ar fi bine să obțin asta și asta". Și ce se întâmplă când majoritatea obiectivelor noastre sunt executate automat? Nu vom avea aceleași motivații pe care le avem astăzi. O întrebare la care mi-aș dori un răspuns este: ce vor alege să facă cu ele însele derivatele oamenilor din viitor? Unul dintre potențialele rezultate negative este să joace jocuri video tot timpul.

Termenul „inteligență artificială" evoluează în utilizarea sa în limbajul tehnic. În prezent, inteligența artificială este foarte populară și oamenii înțeleg oarecum ce înseamnă. Atunci când computerele abia fuseseră inventate, în anii 1940 și 1950, titlul tipic dintr-o carte sau dintr-un articol de ziar despre computere era „creiere electronice uriașe". Ideea era că, la fel cum buldozerele și motoarele cu aburi au automatizat munca mecanică, computerele vor automatiza munca intelectuală. Promisiunea s-a dovedit a fi mai greu de ținut decât se așteptau mulți oameni. La început, exista mult optimism; mulți bani din trezoreria de stat erau cheltuiți pe astfel de eforturi la începutul anilor 1960. Pur și simplu, nu funcționau.

Există multe portretizări cu iz science-fiction ale computerelor din vremea aceea. Unul simpatic este *Desk Set*, despre un computer de tip IBM instalat într-o companie de difuzare a presei și care îi lăsa pe toți fără locuri de muncă. Este simpatic,

deoarece computerului i se pun câteva întrebări de bibliotecă. Când eu și colegii mei construiam Wolfram|Alpha, una dintre ideile pe care le-am avut a fost să-l facem capabil să răspundă la toate acele întrebări de bibliotecă din *Desk Set*. În 2009, putea deja să răspundă la toate.

În 1943, Warren McCulloch și Walter Pitts au introdus un model pentru cum ar putea funcționa creierele conceptual și formal – o rețea neuronală artificială. Au văzut că modelul lor bazat pe creier ar face calcule la fel ca o mașină Turing. Din activitatea lor a reieșit că putem construi rețele neuronale asemănătoare creierului care ar acționa ca niște computere generale. Și, de fapt, activitatea practică desfășurată de angajații ENIAC, John von Newmann și alții pe calculatoare a venit pe linie directă nu din mașinile Turing, ci prin această cotitură a rețelelor neuronale.

Dar rețelele neuronale simple nu au realizat multe. Frank Rosenblatt a inventat un dispozitiv de învățare pe care l-a numit „perceptron", care era o rețea neuronală cu un singur strat. La sfârșitul anilor 1960, Marvin Minsky și Seymour Papert au scris o carte intitulată *Perceptrons*, în care dovedeau practic că perceptronii nu ar putea face nimic interesant, ceea ce este corect. Perceptronii puteau să facă doar distincții liniare între lucruri. Prin urmare, ideea a fost mai mult sau mai puțin abandonată. Oamenii spuneau: „Acești băieți au scris că rețelele neuronale nu pot face nimic interesant, deci nicio rețea neuronală nu poate face nimic interesant, așa că haideți să lăsăm baltă rețelele neuronale". Această atitudine a rezistat ceva vreme.

Între timp, au mai apărut și alte abordări ale inteligenței artificiale. Una era bazată pe o înțelegere, la nivel formal, simbolic, a felului în care funcționează lumea; alta era bazată pe statistici și probabilități. Cu privire la inteligența artificială simbolică, unul dintre cazurile-test a fost: putem învăța un computer să rezolve integrale? Putem învăța un calculator să rezolve operații aritmetice? Existau sarcini de traducere automată, despre care oamenii credeau că ar fi un bun exemplu pentru ce pot face computerele. Ideea este că, până în anii 1970, acea abordare a picat.

Apoi a existat o tendință către dispozitive numite sisteme expert, care au luat naștere la sfârșitul anilor 1970 și începutul anilor 1980. Ideea era ca o mașină să învețe regulile pe care le folosește un expert, dându-și seama astfel ce are de făcut. Și această idee a pierit încet-încet. Apoi inteligența artificială a devenit ceva neserios de urmărit.

Am fost interesat de cum se poate construi o mașină asemănătoare inteligenței artificiale de când am fost copil. Eram interesat în mod special de modul în care se pot lua cunoștințele pe care le-am acumulat noi, oamenii, în civilizația noastră și se poate automatiza răspunsul la întrebări în baza acelor cunoștințe. M-am gândit cum s-ar putea realiza asta simbolic, construind un sistem care ar putea defalca întrebările în unități simbolice și răspunde la ele. Lucram la rețele neuronale în vremea aceea și nu am făcut progrese prea mari, așa că m-am concentrat pe altceva.

În anii 2002–2003, m-am gândit din nou la întrebare: ce ar implica un sistem computațional de cunoștințe? Munca pe care o făcusem până atunci mi-ar arătat că greșeam complet în ipoteza mea inițială despre cum aș putea obține acest rezultat. Ipoteza mea inițială fusese că, pentru a putea face un sistem computațional de cunoștințe serios, mai întâi trebuie să construiești un dispozitiv asemănător unui creier, în care să încarci apoi cunoștințe – la fel cum învață oamenii în educația standard. Acum mi-am dat seama că nu era o linie clară între ce este inteligent și ce este pur și simplu computațional.

Presupusesem că există un mecanism magic care ne făcea mult mai capabili decât orice este pur computațional. Dar această presupunere era greșită. Această realizare a condus la Wolfram|Alpha. Ce am descoperit era că poți să iei o colecție mare de cunoștințe despre lume și poți răspunde automat în baza ei, utilizând doar tehnici computaționale. Era o metodă alternativă de inginerie – o metodă mult mai analogă cu ceea ce face biologia în evoluție.

În concluzie, ce faci de obicei când construiești un program este să-l construiești pas cu pas. Dar poți și să explorezi universul

computațional și să minezi tehnologia din acest univers. De obicei provocarea este aceeași ca în cazul mineritului fizic: adică, să zicem că găsești un depozit de fier, cobalt sau gadoliniu, cu proprietăți magnetice speciale, și folosești această capacitate specială în vederea obținerii unui scop uman, pentru ceva ce vrei tu să facă tehnologia. În cazul materialelor magnetice, există multe căi de a face asta. În ceea ce privește programele, aceeași poveste. Există tot felul de programe, chiar și programe mărunte care fac lucruri complicate. Le-am putea pune în slujba unui scop uman util?

Și cum faci inteligențele artificiale să-ți execute obiectivele? Un răspuns ar fi să vorbești cu ele în limba naturală a omului. Funcționează destul de bine când vorbești cu Siri. Dar, când vrei să spui ceva mai lung sau mai complicat, nu mai funcționează la fel de bine. Avem nevoie de un limbaj de programare care poate reprezenta concepte sofisticate într-un mod care poate fi construit progresiv și nu este posibil în limba naturală. Compania mea a petrecut mult timp încercând să construiască un limbaj bazat pe cunoștințe care încorporează cunoașterea lumii direct în limbaj. Abordarea tradițională de creare a unui limbaj de programare este să faci un limbaj ce reprezintă operații pe care computerele știu intrinsec cum să le execute: alocarea memoriei, atribuirea unor valori variabilelor, iterația, modificarea contoarelor de numere și așa mai departe. Esențialmente, le spui computerelor să facă ceva în condițiile stabilite de tine. Abordarea mea presupunea crearea unui limbaj care servește oamenilor, și nu computerelor, și care să preia lucrurile la care se gândește omul și să le convertească într-o formă pe care computerul o poate înțelege. Am putea încapsula cunoștințele pe care le-am acumulat, atât în ceea ce privește științele, cât și colectarea datelor, într-un limbaj pe care l-am putea utiliza ca să comunicăm cu computerele? Aceasta este marea realizare a ultimilor treizeci și ceva de ani din cariera mea.

În anii 1960, oamenii spuneau lucruri precum „când vom face cutare și cutare, vom ști că avem inteligență artificială"; „când vom rezolva o integrală dintr-un curs de analiză matematică, vom ști că avem inteligență artificială"; „când vom putea avea o

conversație cu un computer și o vom face să pară umană..." etc. Dificultatea consta în „stai așa, computerul nu știe suficiente lucruri despre lume". Întrebai computerul ce zi din săptămână este și probabil putea să-ți răspundă. Îl întrebai cine este președinte și probabil nu-ți putea spune. În acel moment, știai că vorbești cu un computer, și nu cu o persoană. Dar acum, când vine vorba despre aceste teste Turing, oamenii care au încercat să conecteze, de exemplu, Wolfram|Alpha la boții lor de testare Turing observă că boții lor pierd de fiecare dată. Pentru că tot ce trebuie să faci este să începi să-i pui mașinii întrebări sofisticate și îți va răspunde la ele! Niciun om nu poate face asta. După câteva întrebări din domenii diferite, niciun om nu va ști răspunsul la toate acele întrebări, însă sistemul va ști. În acest sens, am obținut deja o inteligență artificială bună, la acel nivel.

Apoi mai sunt sarcini care sunt ușoare pentru oameni, dar tradițional foarte grele pentru mașini. Sarcina standard este identificarea vizuală a obiectelor: ce este acest obiect? Oamenii îl pot recunoaște și-l pot descrie simplu, dar computerul nu avea nicio șansă. Însă acum doi ani am scos un mic sistem de identificare a imaginilor și multe alte companii au făcut ceva similar – al nostru se întâmplă să fie puțin mai bun decât celelalte. Îi arăți o imagine și, pentru aproximativ 10 000 de tipuri de lucruri, îți poate spune ce este. Este distractiv să-i arăți o pictură abstractă și să vezi ce spune. Dar face o treabă destul de bună.

Funcționează cu aceeași tehnologie de rețea neuronală pe care și-au imaginat-o McCulloch și Pitts în 1943 și mulți dintre noi au lucrat la ea la începutul anilor 1980. În anii 1980, oamenii au reușit să facă OCR – recunoașterea optică a caracterelor. Au luat cele douăzeci și șase de litere ale alfabetului și au zis: „Ia să vedem, ăsta este un A? Este un B? Este un C?" și așa mai departe. Acest lucru poate fi realizat pentru douăzeci și șase de posibilități diferite, dar nu poate fi făcut pentru 10 000. Problema era cum scalăm întregul sistem care face acest lucru posibil în prezent. Există probabil cinci mii de substantive comune care pot fi transpuse în imagini în limba engleză, zece mii dacă includem și plante

speciale și insecte pe care oamenii le-ar recunoaște cu o anumită frecvență. Ce am făcut a fost să ne antrenăm sistemul cu treizeci de milioane de imagini care reprezintă toate aceste obiecte. Este o rețea neuronală mare, complicată și haotică. Probabil că detaliile rețelei nu contează, dar sunt necesare un cvadrilion de operații ale unității de procesare grafică pentru această instruire.

Sistemul nostru este impresionant, pentru că aproape că poate să facă ce face un om. Are aproximativ același volum de informații pe care-l folosesc și oamenii – un copil vede aproximativ același număr de imagini în primii doi ani ai vieții sale. Aproximativ același număr de operații trebuie făcute și în procesul de învățare, utilizând aproximativ același număr de neuroni, cel puțin în primul nivel al cortexului nostru vizual. Detaliile sunt diferite; modul în care funcționează acești neuroni artificiali nu prea are de-a face cu modul în care funcționează neuronii creierului. Dar conceptul este similar și există o anumită universalitate pentru ce se întâmplă. La nivel matematic, este o compoziție cu foarte multe funcții, cu anumite proprietăți de continuitate care permit utilizarea anumitor metode de calcul ce permit instruirea incrementală a sistemului. Date fiind aceste atribute, poți obține ceva ce face ce fac și creierele omenești în recunoașterea psihologică.

Dar constituie asta inteligență artificială? Sunt câteva componente de bază. Este vorba despre recunoașterea psihologică, transformarea vocii în text, traducerea limbii – lucruri pe care oamenii reușesc să le facă cu diverse grade de dificultate. În principiu, acestea sunt câteva dintre elementele prin care facem mașinile să pară omenești în ceea ce fac. Pentru mine, unul dintre lucrurile interesante a fost să încorporez aceste capacități într-un limbaj precis și simbolic pentru a reprezenta lumea de zi cu zi. Acum avem un sistem care poate spune „acesta este un pahar cu apă". Putem trece de la imaginea unui pahar cu apă la conceptul paharului cu apă. Acum trebuie să inventăm un limbaj simbolic efectiv pentru a reprezenta aceste concepte.

Am început prin a încerca să reprezint matematic cunoștințe tehnice și am continuat cu alte tipuri de cunoștințe. Am făcut o

treabă foarte bună cu reprezentarea cunoștințelor obiective din lume. Acum problema este reprezentarea discursului uman de zi cu zi într-un mod precis și simbolic – un limbaj bazat pe cunoștințe care are ca scop comunicarea între oameni și mașini, astfel încât oamenii să-l poată citi și mașinile să-l poată înțelege. De exemplu, putem spune „x este mai mare decât 5". Acesta este un predicat. Ai putea spune și „vreau o bucată de ciocolată". Și acesta este un predicat. Are un „vreau" în el. Trebuie să găsim o reprezentare simbolică precisă pentru dorințele pe care le exprimăm în limba naturală omenească.

La sfârșitul anilor 1600, Gottfried Leibniz, John Wilkins și alții se preocupau de ceea ce numeau ei limbajele filosofice, adică reprezentări complete, universale și simbolice ale lucrurilor din lume. Poți analiza limbajul filosofic al lui John Wilkins și poți vedea cum a împărțit ce era important în lume la vremea aceea. Unele aspecte ale condiției umane au fost aceleași încă din anii 1600. Altele sunt foarte diferite. Secțiunea sa despre moarte și diversele forme ale suferinței umane era imensă; în ontologia de astăzi este mult mai mică. Este interesant să vezi cum un limbaj filosofic de astăzi diferă de un limbaj filosofic de la jumătatea secolului al XVII-lea. Este o măsură a progresului nostru. Multe astfel de încercări de formalizare au avut loc de-a lungul anilor. În matematică, de exemplu, *Principia Mathematica* a lui Whitehead și Russell din 1910 a reprezentat cea mai mare paradă de erudiție. Au existat și încercări înaintea lor din partea lui Gottlob Frege și Giuseppe Peano, care au fost puțin mai modeste în prezentare. În cele din urmă, s-au înșelat în ceea ce credeau ei că ar trebui formalizat: considerau că trebuie să formalizeze un proces de dovezi matematice, care se pare că nu este lucrul de care sunt interesați cei mai mulți oameni.

Cu privire la un corespondent modern al testului Turing, este o întrebare interesantă. Există încă botul conversațional, care este ideea lui Turing. Acesta nu a fost rezolvat încă. Va fi rezolvat – singura întrebare este „care este aplicația pentru care este rezolvat?" Mult timp aș fi întrebat „de ce să ne pese?" – pentru că am

crezut că aplicația principală ar fi în domeniul asistenței clienților, care nu era prea sus pe lista mea. Dar asistența clienților, unde încerci să creezi o interfață, este fix locul în care ai nevoie de acest limbaj conversațional.

O diferență mare între vremea lui Turing și a noastră este metoda de comunicare cu computerele. Pe vremea lui, introduceai ceva în mașină și mașina îți afișa un răspuns pe ecran. În ziua de astăzi, răspunde cu un ecran – de exemplu, atunci când vrei să cumperi un bilet la film. Cum este diferită tranzacția făcută cu o mașină față de tranzacția făcută cu un om? Răspunsul principal este că există un afișaj vizual. Te întreabă ceva și tu apeși un buton, putând vedea rezultatul imediat. De exemplu, în Wolfram|Alpha, când este utilizat de Siri, dacă răspunsul este scurt, Siri îți va spune răspunsul scurt. Dar cei mai mulți oameni vor afișajul vizual, care arată informațiile grafice pentru cutare și cutare. Aceasta este o formă de comunicare nonumană care se dovedește a fi mai bogată decât comunicarea umană tradițională rostită sau tastată. În majoritatea comunicărilor de la om la om, suntem limitați la limbajul pur, pe când în comunicarea dintre computer și om avem acest canal de comunicare mult mai larg – comunicarea vizuală.

Multe dintre cele mai puternice aplicații ale testului Turing nu mai sunt relevante acum, pentru că avem acest canal suplimentar de comunicare. De exemplu, iată unul pe care-l urmărim în prezent. Este un bot care comunică pe tema scrierii de cod: îi spui „vreau să scriu un program. Vreau să facă asta". Botul va spune „am scris această bucată de cod. Face asta. Este ceea ce ai vrut?" Bla, bla, bla. Este un bot cu du-te-vino. Dezvoltarea unor astfel de sisteme este o problemă interesantă, pentru că trebuie să aibă un model pentru om dacă încearcă să-ți explice ceva. Trebuie să știe ce nu înțelege omul.

Ce mi-a fost mie greu să înțeleg este: care este scopul unui test Turing? Care este motivația? Ca jucărie, se poate face un mic bot de chat cu care oamenii să discute. Aceasta ar fi următoarea mare descoperire. Runda curentă de învățare profundă – în special rețelele neuronale recurente – face modele destul de bune de

vorbire și scriere umane. Putem să tastăm, de exemplu: „Cum te simți astăzi?" și știe aproape tot timpul ce răspuns să-ți dea. Dar vreau să-mi dau seama dacă pot să-l fac să răspundă automat la e-mailurile pe care le primesc. Știu că răspunsul este nu. Pentru mine, un test Turing bun va fi atunci când un bot ar putea să răspundă la majoritatea e-mailurilor pe care le primesc. Acesta este un test dificil. Va trebui să învețe aceste răspunsuri de la omul la care este conectat e-mailul. Aș putea avea un mic avantaj, pentru că am colectat informații despre mine în ultimii douăzeci și cinci de ani. Am toate e-mailurile primite în ultimii douăzeci și cinci de ani, istoricul tuturor tastelor apăsate de douăzeci de ani. Ar trebui să pot să educ un avatar, o inteligență artificială, care să facă ce pot face eu – poate chiar mai bine decât mine.

Oamenii își fac griji gândindu-se la scenariul în care inteligențele artificiale preiau controlul. Cred că înainte de asta se va întâmpla ceva mult mai amuzant, într-un anumit sens. Inteligența artificială va ști ce intenționezi și se va pricepe să-și dea seama cum să obțină ce dorești. Îi spun GPS-ului mașinii mele că vreau să ajung la o anumită destinație. Nu știu unde naiba mă aflu, doar îmi urmez GPS-ul. Copiilor mei le place să-mi aducă aminte că, odată, când aveam un GPS foarte vechi – tipul care îți spunea „Faceți prima la stânga, prima la dreapta etc." –, am ajuns pe unul dintre pontoanele care coborau în portul din Boston.

Mai important este faptul că va exista o inteligență artificială care îți cunoaște istoricul și știe că, atunci când comanzi cina online, probabil vei vrea aia și aia, sau atunci când îi trimiți un e-mail unei anumite persoane, probabil vei vorbi cu ea despre cutare și cutare. Din ce în ce mai pregnant, inteligențele artificiale ne vor sugera ce ar trebui să facem și bănuiesc că, de cele mai multe ori, oamenii vor fi de acord. Este un sfat bun – mai bine decât la ce te-ai fi gândit tu.

Cât privește scenariul în care acestea preiau controlul, poți face rău cu tehnologia și poți face bine cu ea. Unii oameni vor încerca să facă lucruri teribile cu ea, iar alții, lucruri bune. Unul

dintre lucrurile care-mi plac la tehnologia de astăzi este egalizarea pe care a produs-o. Obișnuiam să mă mândresc cu faptul că am un computer mai bun decât toate cunoștințele mele; acum avem cu toții aceleași computere. Avem aceleași telefoane și aproximativ aceeași tehnologie poate fi folosită de un procent important dintre cei 7 miliarde de oameni de pe planetă. Tehnologia regelui nu mai este diferită de a tuturor celorlalți. Acesta este un avans important.

Marea frontieră acum cinci sute de ani era alfabetizarea. Astăzi este programarea. Programarea așa cum se face în prezent va deveni depășită în scurt timp. De exemplu, oamenii nu mai învață limbaje de programare, deoarece computerele se pricep mai bine să scrie programe decât oamenii și doar un număr mic de oameni trebuie să cunoască detaliile despre cum limbajul comun este compilat în limbajele de programare. Foarte mult din ceea ce fac armate de programatori din zilele noastre este plictisitor de similar. Nu mai există niciun motiv pertinent pentru care oamenii să scrie cod Java sau JavaScript. Putem automatiza procesul de programare astfel încât ce este important merge de la ce vrea omul la execuția de către mașină a dorinței sale, cât mai automatizat cu putință. Astfel este crescută egalizarea, ceea ce mă interesează pe mine. În trecut, dacă doreai să scrii o bucată serioasă de cod sau program pentru ceva important și real, era mult de muncă. Trebuia să știi multe despre programare, trebuia să investești luni de zile, trebuia să angajezi programatori care știau ce fac sau trebuia să înveți singur. Era o investiție mare.

În prezent, nu mai este cazul. O piesă de cod de o linie face deja ceva interesant și util. Permite unui număr mare de oameni care nu puteau să-și determine computerele să facă ceva pentru ei acum să o poată face. Aș vrea să văd mulți copii din jurul lumii învățând noile capacități ale programării bazate pe cunoștințe și apoi să producă coduri care sunt efectiv la fel de sofisticate asemenea celor produse de experți. Acest lucru este realizabil. Suntem în punctul în care oricine poate învăța să facă programare bazată pe cunoștințe și, mai important, poate învăța să gândească computațional.

Mecanica efectivă a programării este ușoară acum. Ce este dificil este să-ți imaginezi ceva într-o manieră computațională.

Cum predai gândirea computațională? În ceea ce privește programarea, este o întrebare interesantă. Să ne gândim la nanotehnologie. Cum am obținut nanotehnologia? Am luat tehnologia așa cum o înțelegem la scară mare și am făcut-o foarte mică. Cum aducem un cip CPU la scară atomică? Fundamental, folosim aceeași arhitectură ca în cazul cipului CPU pe care-l știm și-l iubim cu toții. Aceasta nu este singura abordare pe care o avem. Observând ce pot face programele simple, este posibil să transformăm componente simple și învechite și, cu compilatorul potrivit, să facem din ele ceva interesant. Nu facem calcule la scară moleculară încă, pentru că tehnologia de care dispunem ar putea să producă ce dorim în zece ani, probabil. Dar avem componentele necesare pentru a construi un computer universal. Poate nu știm cum să programăm cu aceste componente, dar, căutând în spațiul programelor posibile, vom începe să colectăm modulele, apoi putem crea un compilator pentru ele. Surprinzător este că lucrurile vechi pot fi transformate în ceva sofisticat, iar etapa de compilare nu este atât de înfiorătoare pe cât credem.

Simpla căutare prin universul computațional și încercarea de a găsi programe – module – care sunt interesante este o abordare bună. O abordare inginerească mai tradițională – încercarea de a construi un computer universal doar prin gândire pură – este o treabă dificilă. Asta nu înseamnă că nu se poate face, dar eu cred că vom putea face lucruri minunate doar căutând componentele și programele posibile pe care le putem face cu ele. Apoi ne întoarcem la problema conectării scopurilor omenești la ce ne oferă sistemul.

O întrebare al cărei răspuns mă interesează este: „Cum va arăta lumea când majoritatea oamenilor vor putea să scrie cod?" Am avut o tranziție, poate acum cinci sute de ani, când doar scribii și un procent mic din populație puteau să citească și să scrie limba naturală. Astăzi un procent mic din populație poate să scrie cod. Cea mai mare parte din codul scris este doar pentru computere.

Nu înțelegi lucrurile dacă citești codul. Dar va veni vremea când, ca urmare a lucrurilor pe care am încercat eu să le fac, codul va fi la un nivel suficient de înalt încât să ofere o descriere minimă pentru ceea ce încerci să faci. Va fi o bucată de cod care poate fi înțeleasă de oameni, dar poate fi și executată de mașini.

Codarea este o formă de exprimare, la fel cum a scrie într-o limbă naturală este o formă de exprimare. Pentru mine, bucățile simple de cod sunt poetice – exprimă ideile foarte curat. Există o latură estetică, la fel cum există și în exprimarea într-o limbă naturală. O caracteristică a codului este că poate fi executat imediat; nu este ca scrisul. Când scrii ceva, cineva trebuie să citească ce ai scris și creierul care citește trebuie să absoarbă gândurile care vin de la persoana care a scris. Privește felul în care au fost transmise cunoștințele în istoria lumii. La nivelul zero, o formă de transmitere a cunoștințelor este genetica – adică există un organism și progeniturile sale au aceleași caracteristici pe care le are el. Apoi urmează transmiterea de cunoștințe, care se întâmplă cu lucruri precum recunoașterea psihologică. O creatură nou-născută are o rețea neuronală cu câteva conexiuni aleatorii în ea și, pe măsură ce creatura se mișcă prin lume, începe să recunoască tipuri de obiecte și învață acele cunoștințe.

Apoi urmează nivelul care a reprezentat marea realizare a speciei noastre, care este limba naturală: abilitatea de a reprezenta cunoștințe suficient de abstract încât să le putem comunica de la creier la creier. Este posibil ca limba naturală să fie una dintre cele mai importante invenții ale speciei noastre. Este ceea ce ne-a condus, în multe privințe, la civilizația pe care o avem.

Mai există un nivel și, probabil, într-o bună zi, va avea un nume mai interesant. Cu programarea bazată pe cunoștințe, avem o modalitate de a crea o reprezentare efectivă a lucrurilor reale din lume într-un mod precis și simbolic. Nu doar că poate fi înțeleasă de creiere și poate fi comunicată altor creiere și computere, dar mai este și imediat executabilă.

Limba naturală ne-a dat civilizația, dar programarea bazată pe cunoștințe ce ne va da? Un răspuns greșit este că ne va da

civilizația inteligențelor artificiale. Asta nu vrem să se întâmple, pentru că inteligențele artificiale vor comunica foarte bine una cu cealaltă și noi vom fi scoși din schemă, pentru că nu există nicio limbă intermediară, nicio interfață cu creierele noastre. Unde ne va conduce acest al patrulea nivel de comunicare bazat pe cunoștințe? Dacă ai fi fost Zog în peșteră și ți-ai fi dat seama că începem să folosim limbajul, ți-ai fi putut imagina civilizația în care trăim? Ce ar trebui să ne imaginăm acum?

Această întrebare se leagă de întrebarea cum ar arăta lumea dacă majoritatea oamenilor ar putea să codeze. În mod clar, multe lucruri triviale s-ar schimba, contractele s-ar scrie în cod, rețetele de restaurant s-ar scrie în cod și așa mai departe. Lucruri simple ca acestea s-ar schimba. Dar s-ar schimba și lucruri mult mai profunde. Creșterea nivelului de alfabetizare ne-a dat birocrația, de exemplu, care exista deja, dar a fost accelerată masiv, conferindu-ne o profunzime mai mare pentru sistemele guvernamentale, în bine și în rău. Cum se leagă lumea codului de lumea culturală?

Să ne gândim la educația din licee. Dacă avem gândire computațională, cum ne afectează ea modul în care studiem istoria? Cum ne afectează modul în care studiem limbile, științele sociale și altele? Răspunsul este că are un efect puternic. Imaginează-ți că scrii un eseu. Astăzi, materia primă pentru eseul unui licean tipic este ceva ce a fost scris deja; elevii nu pot, de obicei, să genereze cunoștințe noi cu ușurință. Dar în lumea computațională nu va mai fi cazul. Dacă elevii știu ceva despre cum se scrie un cod, vor accesa toate informațiile istorice digitalizate și vor ajunge să creeze ceva nou. Apoi vor scrie un eseu despre ceva ce au descoperit. Realizarea programării pe bază de cunoștințe se bazează pe faptul că nu mai este sterilă, pentru că are cunoștințele lumii întrețesute în limba pe care o utilizezi pentru a scrie codul.

Computații sunt în tot universul: într-un fluid turbulent care produce un tipar complicat de curgere, în mecanica celestă a interacțiunilor planetare, în creiere. Dar are tehnica de calcul un

scop? Poți pune această întrebare referitoare la orice sistem? Vremea are un scop? Clima are un scop?

Cineva care privește Pământul din spațiu poate să-și dea seama dacă ceva are scop acolo jos? Există o civilizație acolo? În Marele Lac Sărat din Utah se află o linie dreaptă. Se pare că este un dig circulabil care împarte două zone ale lacului ce au alge de culori diferite, deci este o linie dreaptă foarte dramatică. Există un drum în Australia care este lung și drept. Există o cale ferată în Siberia care este lungă și se aprind lumini când trenul oprește într-o gară. Deci, din spațiu, se pot vedea linii drepte și tipare.

Dar sunt acestea exemple suficient de clare de scopuri clare pe Pământ dacă sunt observate din spațiu? De fapt, cum ne-am da seama dacă în spațiu sunt extratereștri? Cum ne dăm seama dacă un semnal pe care-l interceptăm indică un scop? Pulsarii au fost descoperiți în 1967, când am detectat o oscilație periodică la fiecare secundă. Prima întrebare a fost dacă nu cumva este un semnal? Întrucât ce altceva ar emite un semnal periodic? S-a dovedit a fi o stea neutronică rotativă.

Un criteriu de aplicat unui fenomen care prezintă potențialul de a avea un scop este dacă îndeplinește condițiile acțiunii minime pentru atingerea scopului. Dar înseamnă asta că a fost construit pentru scopul respectiv? Mingea se rostogolește în josul pantei datorită atracției gravitaționale. Sau mingea se rostogolește în josul pantei pentru a satisface principiul acțiunii minime. De obicei, acestea sunt cele două explicații pentru o acțiune care pare că are un scop: explicația mecanicistă și explicația teleologică. Practic, toată tehnologia noastră existentă eșuează la testul acțiunii minime în atingerea scopului său. Cele mai multe lucruri pe care le construim noi sunt înrădăcinate adânc în istoricul tehnologic și sunt incredibil de nonminime în ceea ce privește atingerea scopului. Gândește-te la un cip CPU; nu-i nicio șansă ca aceea să fie calea minimă de a obține ce obține un cip CPU.

Această problemă a identificării discernământului este una grea. Este o întrebare importantă, pentru că zgomotele radio venite din galaxie sunt foarte similare transmisiunilor CDMA ale

telefoanelor mobile. Aceste transmisii folosesc secvențe de pseudozgomot, care se întâmplă să aibă anumite proprietăți de repetabilitate. Dar sunt percepute ca zgomot și sunt concepute ca zgomot pentru a nu interfera cu alte canale. Problema devine și mai neclară. Dacă am analiza o secvență de numere prime generate de un pulsar, ne-am întreba ce le-a generat. Ar însemna că o întreagă civilizație a crescut, a descoperit numerele prime și a inventat computerele și transmițătoarele radio, după care a făcut această emisie? Sau este vorba doar despre un proces fizic care emite numere prime? Există un automat celular care emite numere prime. Poți vedea cum funcționează dacă-l dezasamblezi. Are ceva mic ce ricoșează în el și din acest mecanism reiese o secvență de numere prime. Nu a avut nevoie de întreaga istorie a civilizației și a biologiei pentru a ajunge în acest punct.

Nu cred că există un „scop" abstract, *per se*. Nu cred că există sensuri abstracte. Universul are un scop? Cu această întrebare eșuezi în teologie. Nu există niciun sens important în care să existe o noțiune abstractă a scopului. Scopul este ceva ce reiese din istorie.

Unul dintre lucrurile care ar putea fi adevărate despre lumea noastră este că poate trecem prin toată istoria, biologia și civilizația noastră și, după toate acestea, răspunsul este „42" sau ceva de genul acesta. Am trecut prin patru miliarde de ani de diverse tipuri de evoluție și rezultatul este „42".

Nu se va întâmpla nimic de acest gen, datorită ireductibilității computaționale. Există procese computaționale pe care le putem parcurge în care nu există nicio scurtătură pentru proces. O mare parte din știință se preocupă de găsirea unei scurtături pentru computațiile făcute de natură. De exemplu, dacă facem mecanică celestă și vrem să prezicem poziția planetelor de peste un milion de ani, am putea urma ecuațiile pas cu pas. Dar marea realizare a științei este că putem să scurtăm acest proces și să reducem calculele. Putem fi mai inteligenți decât Universul și-i putem prezice punctul final fără a trece prin toate etapele. Dar chiar și cu o mașină suficient de inteligentă și cu o matematică suficient de

bună, nu putem ajunge la punctul final fără a trece prin etape. Unele detalii sunt ireductibile. Trebuie să urmăm ireductibil acele etape. Acesta este motivul pentru care istoria înseamnă ceva. Dacă am putea ajunge la punctul final fără a trece prin etape, istoria ar fi, într-un anumit sens, inutilă.

Așadar ideea nu este că noi suntem inteligenți și toate celelalte din lume nu sunt. Nu este nicio diferență abstractă enormă între noi și nori, și automatele celulare. Nu putem spune că această rețea neuronală asemănătoare creierului este calitativ diferită de acest sistem automat al celularului. Diferența este o diferență detaliată. Această rețea neuronală asemănătoare creierului a fost produsă de istoria lungă a civilizației, în timp ce automatul celular a fost creat de computerul meu în ultima microsecundă.

Problema inteligenței artificiale abstracte este similară problemei recunoașterii inteligenței extraterestre: cum stabilești dacă are sau nu un scop? Aceasta este o întrebare la care eu consider că nu s-a dat un răspuns. Spunem ceva de genul „Ei bine, inteligența artificială va fi inteligentă când va putea face cutare și cutare". Când va putea găsi numere prime. Când va putea produce cutare și cutare obiect. Dar există multe alte căi de a obține acele rezultate. Din nou, nu există nicio linie despărțitoare clară între inteligență și simpla tehnică de calcul.

Este o altă parte a poveștii copernicane: obișnuiam să credem că Pământul este centrul Universului. Acum credem că suntem speciali pentru că avem inteligență și nimic altceva nu o mai are. Mă tem că vestea proastă ar putea fi că aceasta nu este o distincție.

Iată unul dintre scenariile mele. Să zicem că va veni o vreme în care conștiința umană se va putea încărca ușor pe o platformă digitală, virtualizată și cu toate cele, și, destul de repede, am avea o cutie cu un trilion de suflete. În cutie se află un trilion de suflete, toate virtualizate. În cutie, computațiile neîntrerupte vor fi moleculare – poate derivate din biologie, poate nu. Dar cutia va face tot soiul de lucruri complicate. Lângă cutie se află o piatră. În piatră, se întâmplă tot timpul lucruri complicate, tot felul de particule subatomice care fac tot felul de lucruri. Care este

diferența dintre piatră și cutia cu un trilion de suflete? Răspunsul este că detaliile legate de ceea ce se întâmplă în cutie au fost derivate dintr-o istorie lungă a civilizației umane, inclusiv ce au văzut oamenii pe YouTube cu o zi înainte. În timp ce piatra are istoria ei geologică lungă, dar nu și istoria civilizației noastre.

Realizarea că nu există o distincție veritabilă între inteligență și simpla tehnică de calcul te face să-ți imaginezi acest viitor – punctul final al civilizației noastre ca o cutie cu un trilion de suflete, fiecare dintre ele jucând, practic, un joc video, pentru totdeauna. Care este „scopul" aici?

Inteligența artificială este povestea zilei de astăzi, povestea din spatele tuturor celorlalte povești, este a doua venire a lui Iisus și Apocalipsa în același timp: inteligența artificială bună și inteligența artificială malefică. Această carte a reieșit în urma unor conversații în desfășurare cu câțiva gânditori importanți, atât din lumea inteligenței artificiale, cât și din afara acesteia, despre ce este și ce înseamnă inteligența artificială.

Noi credem că, dacă urci pe raftul cărților citite, ai o perspectivă din ce în ce mai largă asupra lumii în care trăim. Pentru Versant am ales cu pasiune, responsabilitate și discernământ acele cărți care ne inspiră și ne dezvăluie universuri nebănuite. Versant înseamnă nonfiction revelator și incitant. Vă invităm, ca împreună, să descoperim cât de sus ne poate duce.

„Douăzeci și cinci de oameni au scris eseuri, fiecare în parte fiind îngrijorat de ce se întâmplă astăzi în era inteligenței artificiale. Minți posibile nu este cartea mea, mai degrabă este cartea noastră: Seth Lloyd, Judea Pearl, Stuart Russell, George Dyson, Daniel C. Dennett, Rodney Brooks, Frank Wilczek, Max Tegmark, Jaan Tallinn, Steven Pinker, David Deutsch, Tom Griffiths, Anca Drăgan, Chris Anderson, David Kaiser, Neil Gershenfeld, W. Daniel Hillis, Venki Ramakrishnan, Alex „Sandy" Pentland, Hans Ulrich Obrist, Alison Gopnik, Peter Galison, George M. Church, Caroline A. Jones și Stephen Wolfram."

John Brockman

„Dacă vrem ca inteligențele artificiale foarte capabile să fie compatibile cu oamenii, nu le putem crea izolate de oameni, după care să încercăm să le facem compatibile; mai degrabă, va trebui să definim inteligența artificială <<compatibilă cu omul>> încă de la început."

Anca Drăgan, Departamentul de Inginerie Electrică și Informatică de la UC Berkeley

„Deși sunt oficial un libertarian civil, profund preocupat de amenințările contemporane la adresa exprimării libere, nu mă înspăimântă avansurile tehnologice ale internetului, tehnologiei video sau inteligenței artificiale."

Steven Pinker, psiholog, Universitatea Harvard

VERSANT

#popscience
#inteligențăartificială
#eseistică

www.vellant.ro
9 786069 800676